FACULTY OF ENGINEERING
AND ARCHITECTURE

Co-Speech Gestures: Evaluation and Generation

Andries Pieter Wolfert

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Computer Science Engineering

Supervisors
Prof. Tony Belpaeme, PhD - Prof. Francis wyffels, PhD
Department of Electronics and Information Systems
Faculty of Engineering and Architecture, Ghent University

July 2024

GHENT
UNIVERSITY

ISBN 978-94-6355-856-3
NUR 984
Wettelijk depot: D/2024/10.500/61

Members of the Examination Board

Chair

Prof. Luc Dupré, PhD, Ghent University

Other members entitled to vote

Prof. Veronique Hoste, PhD, Ghent University
Prof. Panos Markopoulos, PhD, Technische Universiteit Eindhoven, the Netherlands
Prof. Jelle Saldien, PhD, Ghent University
Prof. Paul Vogt, PhD, Rijksuniversiteit Groningen, the Netherlands

Supervisors

Prof. Tony Belpaeme, PhD, Ghent University
Prof. Francis wyffels, PhD, Ghent University

Voor Anouk en Lua

Dankwoord

Na ruim vijf jaar werken aan mijn doctoraat, en vertoeven in België (met wat speeltijd in Zweden en Portugal) mag ik eindelijk het dankwoord schrijven van mijn proefschrift. Alhoewel, ik schrijf wel mijn proefschrift, maar dit proefschrift was er nooit geweest zonder de steun en hulp van vele mensen. Mijn dank is groot aan allen die een rol hebben gespeeld in mijn persoonlijke en professionele ontwikkeling, die mij ertoe in staat hebben gesteld het werk voor dit proefschrift te voltooien.

Allereerst wil ik mijn grote dank en waardering uitspreken aan Tony. Wat begon met een sollicitatiegesprek van 30 minuten (welke veel langer duurde), eindigde in een aanbod voor een promotieplek. Je gaf de juiste tegendruk wanneer het nodig was, maar liet mij ook gaan toen ik daar nood aan had. De vrijheid en ruimte die je gaf zorgde ervoor dat ik mij tot een volwaardig onderzoeker kon ontwikkelen, en je hebt mij geïnspireerd om te netwerken, en om samen met anderen wetenschap te bedrijven. Ik zal nooit vergeten dat ik dankzij jou mijn thesis heb kunnen schrijven onder de warmte van de Portugese zon op 2 minuten stappen van de zee.

Maxim en Mathieu, we zaten in hetzelfde schuitje al die jaren, en dat heeft ons verbroederd. Veel dank voor de vele wandelingen, gezaag over het weer, avondjes op café, frieten en jullie vriendschap.

Ik wil al mijn collega's in Gent, over de jaren heen, bedanken voor hun input en gezelligheid. Bedankt Alexander, Andreas, Annelies, Asma, Axel, Benedikt, Bjarne, Dries, Eva, Francis, Gabriel, Giulio, Ira, Jarne, Jeanne, Joni, Len, Luthffi, Maria, Marieke, Matthias, Matthijs, Natacha, Olivier, Peter, Pieter, Qiaoqiao, Raveena, Rembert, Remko, Ruben, Saya, Stefan, Tanguy, Thomas, Tom, Victor, Yi en Zimcke.

I am incredibly grateful to my colleagues and friends at KTH in Stockholm, Sweden. What started with a presentation in August 2019 at the KTH campus, resulted in a four month stay in 2022. Gustav, I cherish our long discussions we have had about our research. Taras, your friendliness and your ambition for doing research have been an inspiration ever since. Jonas and Joakim and all others at TMH, thank you for your input and friendly welcome. I truly felt at home during my stay at KTH. Tack så mycket!

I would like to thank Nicole, Patrik, Jeffrey, Youngwoo, Teodor, Carla, and Mihail. Thanks to all kind researchers and friends at GAIPS Tecnico in Lisbon for their warm welcome, as a friend of a friend I enjoyed spending time with you.

Sofia, my star twin, thank you for all that you have done for me. I enjoyed spending time with you in Stockholm, our long walks and talks, and countless diners. I hope our paths will keep on crossing.

Ezra, wat fijn om een nuchtere Fries als vriend te hebben, te midden van al die Vlamingen in Gent. Wat hebben we genoten van de vele fietstochtjes, de wandelingen, de roadtrip naar Fryslân, liters aan Belgisch bier en de kilo's frieten met mayonaise en samurai saus. Dankjewel voor je heldere en ontnuchterende inzichten die je hebt kunnen geven gedurende al die jaren. Ward, dankjewel voor al die wandelingen in en om Gent, en de ritjes naar Vleteren.

Mart, Koen en Chris, dank jullie wel voor alle steun, en niet alleen op academisch vlak. Jullie advies en bereidheid tot een luisterend oor heeft mij veel rust gegeven. Dank jullie wel Michelle, Tom, Evert, Jacob, Frans, Laura, Rianne, Job, Matthijs, Gabrielle, Silko en Lotte voor alle fijne momenten die we hebben gehad.

Mam, Pap, Eric, Jorieke en Nienke, dank voor jullie steun al die jaren, zonder jullie support was ik het eiland nooit afgekomen.

Allerliefste Anouk, wat eens begon met een idee over samenwerken, leidde uiteindelijk tot de geboorte van onze mooie dochter Lua. Woorden schieten tekort om te beschrijven hoe belangrijk jij voor mij bent in mijn leven, mijn steun en toeverlaat. Ik ben zo ontzettend dankbaar voor je steun, en zo gelukkig dat ik met jou een gezin mag vormen. Ik hoop nog lang iedere dag met een glimlach naast je wakker te worden.

Pieter Wolfert, Juni 2024

Summary

In this thesis we take a look at subjective evaluations for generated co-speech gestures for embodied conversational agents (ECAs), such as on-screen characters and robots and the generation of nonverbal speech and listening behaviour for ECAs. With subjective evaluations we mean the evaluation of stimuli by human participants in online and offline studies.

In our day to day communication with others we rely on both verbal and nonverbal behaviour. The nonverbal component is often referred to as 'everything but words', and covers behaviours such as facial expressions, body pose, gestures and even walking gait. Nonverbal behaviour is of importance for our communication, as it has been shown that for example the use of gestures helps to get a message across; not only for the listener, due to the multimodality of the communication, but also for the speaker, where use of nonverbal behaviour improves the flow of communication. In our encounters with ECAs, we expect them to be 'like' us, and to understand our behaviour. Moreover, equipping embodied conversational agents with the ability to display nonverbal behaviour, such as gestures, turned out to make these ECAs more persuasive, even more when combined with other nonverbal channels.

In the last few years, generating gestures using data driven approaches gained traction, thanks to the steep rise of deep learning and computational power. For the task of generating gestures in ECAs, large deep neural networks are trained on datasets to learn to synthesise new gestures. Often, the synthesis quality is assessed through objective metrics, such as the average velocity or jerk of the motion. Additionally, subjective evaluations are used to test for the naturalness of the motion or the appropriateness of the motion in relation to other (non)verbal signals. However, an extensive review of 22 papers by us of the field showed that the use of subjective evaluation methods varied, both in terms of the depth of the evaluations as well as how results were reported. We also find that many studies do not include demographic information, which is crucial for other researchers to reproduce existing research.

Comparing and Evaluating Gestures

We look at four different methods for comparing generated motion subjectively. First, we consider hand generated and computer generated beat gestures, and use a ranking approach with pairwise comparisons to assess the 'best' condition. Second, we run a comparative study where we look into using rating scales (with scores ranging from 1 to 100) with one direct question versus pairwise comparisons. Not only do we look into the difference of reported scores, but we also take into account the completion time and user comments. Lastly, we evaluate a newly designed questionnaire for assessing the human-likeness, appropriateness and intelligibility of stimuli, based on the identified constructs from our review on subjective evaluation methods, discussed earlier. We find that using a ranking approach only does not provide that much information. This becomes more clear from our comparison study between pairwise comparisons and rating scales; from which we can conclude that rating scales provide more information and are also easier to use with a larger number of conditions. Finally, we had hoped to validate our questionnaire so that we could present an additional subjective instrument for the assessment of computer generated gesticulation stimuli. But, our results were not conclusive, there were no significant differences between the conditions on the constructs of the questionnaire. Which was rather surprising especially since we would have expected differences between the two evaluated systems and recorded human motion.

Crowdsourcing Gesture Generation Systems and Evaluations

We further report on two challenges we co-organized. The first GENEA Challenge was organized at IVA 2020, and the second GENEA Challenge was organized at ICMI 2022. These two challenges had both the task and aim in common: generating co-speech gestures for an ECA. The first challenge relied on a dataset containing a single speaker holding a monologue, and the second challenge relied on multiple speakers and dyadic conversations. In the first challenge the appropriateness and human-likeness results overlap largely, but this is successfully disentangled in the second challenge. Where we used similar evaluation techniques in the first challenge, we opted for a different approach for the second challenge: mismatching stimuli. With the mismatching paradigm, non-matching nonverbal behaviour was combined with audio and paired with (generated) matching behaviour. This approach improved our understanding of evaluating the human-likeness of generated nonverbal behaviour. For the first time we also had a participating team that scored higher on human-likeness than the recorded human motion condition. Fully measuring all aspects of human-likeness of generated motion is a complex task, and our results indicated that we might have reached the limit of the

current methodology. Besides these results, these challenges turn out to have a lasting effect on the standard of reporting in later work, and the willingness to share code and datasets with others in the community.

Generating Nonverbal Behaviour

Our last contribution covers the generation of both speech and listening behaviour. We took our experience from our review, our comparative user studies and challenges, to aim for generating nonverbal behaviour. We used an existing model that was built for single speaker gesture generation, and adapted it so it could generate dyad aware nonverbal listening and speaking behaviour, by leveraging the information of both speakers and their identity. We compared our adapted model to a baseline model from the GENEA Challenge, that won the reproducibility award. First, we tested whether participants could properly identify the listening behaviour segments. Second, we tested human-likeness for gesticulation. Third, we assessed the appropriateness of the gestures for the speech. Fourth, we assessed human-likeness for listening, followed by appropriateness of the motion for listening. In terms of human-likeness, the models did not differ significantly from each other, but did so with the recorded human motion. For listening, a similar pattern unfolded, where human-likeness between the two models was not significantly different. For the appropriateness dimension, we could not report any significant differences between the three conditions compared. Our adaptation of this model seems to be doing what we expect it to be doing.

Throughout this thesis we find that evaluating generated nonverbal behaviour is not an easy task. We compare a large variety of subjective evaluation methods that do provide different information. However, the contributions in this thesis shed light on how to effectively evaluate generated nonverbal behaviour, and the importance of it.

Samenvatting

In dit proefschrift kijken we naar subjectieve evaluaties van gegenereerde co-spraakgebaren voor belichaamde gespreksagenten (ECA's), zoals virtuele personages en robots, en het genereren van non-verbale spraak en luistergedrag voor ECA's. Met subjectieve evaluaties bedoelen we de evaluatie van stimuli door menselijke deelnemers in online en offline onderzoeken.

In onze dagelijkse communicatie met anderen zijn we afhankelijk van zowel verbaal als non-verbaal gedrag. De non-verbale component wordt vaak 'alles behalve woorden' genoemd en omvat gedragingen zoals gezichtsuitdrukkingen, lichaamshouding, gebaren en zelfs ons looppatroon. Non-verbaal gedrag is van belang voor onze communicatie, omdat is aangetoond dat bijvoorbeeld het gebruik van gebaren helpt om een boodschap over te brengen; niet alleen voor de luisteraar, vanwege de multimodaliteit van de communicatie, maar ook voor de spreker, waarbij het gebruik van non-verbaal gedrag de communicatiestroom verbetert. In onze ontmoetingen met ECA's verwachten we dat ze zijn zoals wij, en ons gedrag begrijpen. Bovendien bleek het uitrusten van ECA's met de mogelijkheid om non-verbaal gedrag te vertonen, zoals gebaren, deze ECA's overtuigender te maken, zelfs nog meer in combinatie met andere non-verbale kanalen zoals bijvoorbeeld gezichtsuitdrukkingen.

De afgelopen jaren heeft het genereren van gebaren met behulp van datagestuurde benaderingen aan kracht gewonnen, dankzij de sterke opkomst van deep learning en computer rekenkracht. Voor de taak om gebaren in ECA's te genereren, worden grote diepe neurale netwerken getraind op datasets om nieuwe gebaren te leren synthetiseren. Vaak wordt de synthese kwaliteit beoordeeld aan de hand van objectieve meetgegevens, zoals de gemiddelde snelheid van de beweging. Daarnaast worden subjectieve evaluaties gebruikt om de natuurlijkheid van de beweging of de geschiktheid van de beweging in relatie tot andere (non)verbale signalen te testen. Uit een uitgebreid onderzoek van 22 artikelen door ons uit het veld bleek echter dat het gebruik van subjectieve evaluatiemethoden varieerde, zowel wat betreft de diepgang van de evaluaties als de manier waarop de resultaten werden gerapporteerd. We constateren ook dat veel onderzoeken geen demografische informatie rapporteren, wat cruciaal is voor andere onderzoekers om bestaand onderzoek te reproduceren.

Vergelijking en Evaluatie van Gebaren

We bekijken vier verschillende methoden om gegenereerde bewegingen subjectief te vergelijken. Ten eerste beschouwen we met de hand gegenereerde en computer gegenereerde 'beat' gebaren, en gebruiken we een rangschikking benadering met paarsgewijze vergelijkingen om de 'beste' conditie te beoordelen. Ten tweede voeren we een vergelijkend onderzoek uit waarin we kijken naar het gebruik van beoordelingsschalen (met scores variërend van 1 tot 100) met één directe vraag versus paarsgewijze vergelijkingen. We kijken niet alleen naar het verschil tussen de gerapporteerde scores, maar houden ook rekening met de doorlooptijd en opmerkingen van gebruikers. Ten slotte evalueren we een nieuw ontworpen vragenlijst voor het beoordelen van de menselijke gelijkenis, geschiktheid en begrijpelijkheid van stimuli, gebaseerd op de geïdentificeerde constructies uit ons overzicht van subjectieve evaluatiemethoden, die eerder zijn besproken. Wij vinden dat het gebruik van alleen een rangschikkingsbenadering niet zoveel informatie oplevert. Dit wordt duidelijker uit ons vergelijkingsonderzoek tussen paarsgewijze vergelijkingen en beoordelingsschalen; waaruit we kunnen concluderen dat beoordelingsschalen meer informatie verschaffen en ook gemakkelijker te gebruiken zijn met een groter aantal voorwaarden. Ten slotte hadden we gehoopt onze vragenlijst te valideren, zodat we een aanvullend subjectief instrument konden presenteren voor de beoordeling van door de computer gegenereerde gebarenstimuli. Maar onze resultaten waren niet overtuigend, er waren geen significante verschillen tussen de condities op de constructen van de vragenlijst. Dat was nogal verrassend, vooral omdat we verschillen hadden verwacht tussen de twee geëvalueerde systemen en de geregistreerde menselijke bewegingen.

Crowdsourcing Systemen voor Gebarengeneratie en Evaluaties

Verder doen we verslag van twee uitdagingen die we mede hebben georganiseerd. De eerste GENEA Challenge werd georganiseerd op IVA 2020 en de tweede GENEA Challenge werd georganiseerd op ICMI 2022. Deze twee uitdagingen hadden zowel de taak als het doel gemeen: het genereren van co-speech-gebaren voor een ECA. De eerste uitdaging was gebaseerd op een dataset met daarin één enkele spreker die een monoloog hield, en de tweede uitdaging was gebaseerd op meerdere sprekers en dyadische gesprekken. Bij de eerste uitdaging overlappen de resultaten van geschiktheid en menselijke gelijkenis grotendeels, maar dit wordt met succes ontward in de tweede uitdaging. Waar we bij de eerste uitdaging soortgelijke evaluatietechnieken gebruikten, kozen we voor een andere aanpak voor de tweede uitdaging: het niet matchen van stimuli. Met het mismatching-paradigma werd niet-matchend non-verbaal gedrag gecombineerd met audio en gepaard met (gegenereerd)

matchinggedrag. Deze aanpak verbeterde ons begrip van het evalueren van de menselijke gelijkenis van gegenereerd non-verbaal gedrag. Voor het eerst hadden we ook een deelnemend team dat hoger scoorde op menselijke gelijkenis dan op de geregistreerde menselijke bewegingsconditie. Het volledig meten van alle aspecten van de menselijke gelijkenis van gegenereerde beweging is een complexe taak, en onze resultaten gaven aan dat we mogelijk de limiet van de huidige methodologie hebben bereikt. Naast deze resultaten blijken deze uitdagingen een blijvend effect te hebben op de standaard van rapportage in later werk, en op de bereidheid om code en datasets te delen met anderen in de gemeenschap.

Non-verbaal Gedrag Genereren

Onze laatste bijdrage gaat over het genereren van zowel spraak- als luistergedrag. We hebben onze ervaringen uit onze review, onze vergelijkende gebruikersstudies en uitdagingen gebruikt om te streven naar het genereren van non-verbaal gedrag. We gebruikten een bestaand model dat was gebouwd voor het genereren van gebaren door één spreker, en pasten het aan zodat het dyade-bewust non-verbaal luister- en spreekgedrag kon genereren, door gebruik te maken van de informatie van beide sprekers en hun identiteit. We vergeleken ons aangepaste model met een basismodel van de GENEA Challenge, dat de reproduceerbaarheidsprijs won. Eerst hebben we getest of deelnemers de segmenten van het luistergedrag goed konden identificeren. Ten tweede hebben we de menselijke gelijkenis getest op gebaren. Ten derde beoordeelden we de geschiktheid van de gebaren voor de toespraak. Ten vierde beoordeelden we de menselijke gelijkenis voor het luisteren, gevolgd door de geschiktheid van de beweging voor het luisteren. Qua menselijke gelijkenis verschilden de modellen niet significant van elkaar, maar dat deden ze wel met de vastgelegde menselijke beweging. Bij het luisteren ontvouwde zich een soortgelijk patroon, waarbij de menselijke gelijkenis tussen de twee modellen niet significant verschilde. Voor de dimensie passendheid konden we geen significante verschillen melden tussen de drie vergeleken condities. Onze aanpassing van dit model lijkt te doen wat we ervan verwachten.

In dit proefschrift zien we dat het evalueren van gegenereerd non-verbaal gedrag geen gemakkelijke taak is. We vergelijken een grote verscheidenheid aan subjectieve evaluatiemethoden die wel verschillende informatie opleveren. De bijdragen in dit proefschrift werpen echter licht op hoe effectief gegenereerd non-verbaal gedrag kan worden geëvalueerd, en op het belang ervan.

Contents

List of Figures

List of Tables

Acronyms

AMT Amazon Mechanical Turk

APE Average Position Error

BLT Bradley-Terry Luce

BML Behaviour Markup Language

DNN Deep Neural Network

ECA Embodied Conversational Agent

FGD Fréchet gesture distance

fps frames per second

GENEA Generation and Evaluation of Non-verbal Behaviour for Embodied Agents

IVA International Conference on Intelligent Virtual Agents

MoCap Motion Capture

ms milliseconds

MSE Mean Squared Error

MUSHRA Multiple Stimuli with Hidden Reference and Anchor

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

ROS Robot Operating System

SD Standard Deviation

URDF Unified Robot Description Format

XML Extensible Markup Language

Introduction

1

Introduction

Nowadays, there is an increasing chance that you will interact with a (virtual) agent. This could be through chatbots, on the internet, or in real life with an actual robot. Our initial excitement often fades, since most of these agents we encounter come with fixed behaviours and behavioural loops, and lack personalization. Besides the expectations that we can have full conversations with these agents, we might also expect these agents to understand our body language, and for agents to display body language we can interpret. Over recent years, more and more efforts have been put into generating nonverbal behaviour for ECAs, which has resulted in many approaches and models for the generation of nonverbal behaviour.

Although the concept of open science gains more and more traction, most work does not share code, model weights or datasets, which makes it hard to compare the output of different models built on various datasets. Additionally, numerous approaches for objective evaluations, such as calculating the jerk or average velocity and subjective evaluations, such as pair wise comparisons and the use of rating, have been used over the years to evaluate generated nonverbal behaviour in ECAs, but especially for subjective evaluation methods, there are no standardised approaches. To help move the field forward, we look into a variety of subjective evaluations, their qualities and sensitivities and the benefits of crowd sourcing user evaluations.

1.1 Nonverbal Behaviour for Embodied Conversational Agents

When we interact with an ECA, we encounter two forms of behaviour: verbal and nonverbal behaviour. Where verbal behaviour covers natural language, often in the form of speech, nonverbal behaviour covers aspects such as body language, gestures and facial expressions. Some studies have looked into the total amount of nonverbal behaviour that is part of our behaviour, and reported figures for the the total nonverbal behaviour as part of our communication ranging between 70% and 93%. Mehrabian and Wiener (1967) introduced the 55/38/7 rule, which breaks down communication into three components: 55% of communication is attributed to facial expressions, 38% to tone of voice, and only 7% to the actual words spoken. This conclusion

was based on the scoring of attitudes on videos by participants. However, we use more nonverbal channels than facial expressions alone, and these results should not be generalized to all kinds of face-to-face communication Lapakko, 2015. Our communication is also not only pre-occupied with how we communicate, but also what our intent is for communication. Hence, we should not take these kinds of numbers too serious.

We know of the importance of nonverbal communication in human-human communication, and including nonverbal behaviour in human-agent interaction is of an equal importance. This can be done through rule-based systems, either directly mapping the content of a speech signal to behaviour, or by going from the communicative intent to behaviour. More recent work relies on data-driven generative approaches, keeping in mind the one-to-many nature of nonverbal behaviour, i.e., each input signal could have a different outcome. We are not only interested in one-to-many mappings for single agents, but have an interest in more dyadic aware gesture generation for ECAs. Simply because the way we gesture is not only dependent on what we want to say, or how we want to say it (communicative intent), but also where and with whom we are.

1.2 Objective Assessments

Since nonverbal behaviour can be expressed as a motion signal, it is possible to apply various metrics to calculate the difference in new generated motion to the ground truth recorded motion. One way of doing so is by looking at the derivative of joint positions, known as jerk, and calculate it for both the generated motion for each joint and for the ground truth motion. Through this way it is possible to see how close the generated movement is to the motion captured gesticulation. Additionally, we can look at so called speed histograms, and calculate the hellinger distance between these histograms for different conditions and the recorded motion. Or train an autoencoder on the dataset to compare different conditions in latent space. Finally, we can use a ranking approach once we have calculated objective metrics, and compare rankings for one objective metric with another. We cover objective assessments more in depth in our background section.

1.3 Subjective Assessments

The end goal of most research on gesture generation and evaluation is to equip an agent with the ability to gesture, during an interaction. Since these interactions will be with humans, it is important to bring the human in the loop at various stages of the development. One of the most common ways of doing so, is by involving users through user studies. One could for example decide to show multiple videos rendered from one or more models to users,

and ask these users to rate how natural the gestures are in the video. Or, two videos are presented, and the user has to select the best looking one, or the one best matching the speech. There are several methods of assessing the quality of generated stimuli in a subjective way, and further in this thesis we look at and compare several methods for subjectively evaluating computer generated nonverbal behaviour.

1.4 Research Outline

In this thesis the main focus is on the subjective evaluation of generated co-speech gestures in ECAs. We are concerned with improving the subjective evaluation standards, and applying these improvements ourselves. This topic does not often receive enough attention, and the quality of subjective evaluations in the field has been a mixed bag (at least up till now). Central to this thesis is the question:

"Can we improve and advance the standard of subjective evaluations for the field of nonverbal behaviour generation?"

The setup of this dissertation is as follows: First, in chapter 2 we dive into the field of co-speech gesture generation, and look at how co-speech gestures were generated before the advent of data-driven approaches. We then explore more recent generation strategies, including datasets that have come available. This is followed by a discussion of objective metrics used throughout this work, and a review on subjective evaluations used till now. As evaluation is a key topic to this thesis, we further explore objective and subjective evaluations used up till now. In chapter 3 we look at several evaluation paradigms for nonverbal behaviour. First, we take a look at a ranking approach. Then, we compare rating stimuli to pairwise comparisons, and we finish with a look at the creation and evaluation of a questionnaire, based on constructs that followed our review from chapter 2. For chapter 4, we report on two challenges we co-organised. More specifically, we look at the similarities between the challenges, the results, and synthesise the findings keeping the future of the field in mind. In chapter 5 we generate our own nonverbal behaviour, we do not only look at the generation of dyadic aware speaking behaviour, but we also look at generating listening behaviour. Chapter 6 contains our overall conclusions in relation to the evaluation of nonverbal behaviour, and we provide future perspectives and key ideas that still need to be covered.

1.5 Chapters & Publications

All chapters in this book are based on work published and performed during the course of the PhD. The introduction and discussion chapters are notable

exceptions. Chapter 3 is based on two published works. Chapter 4 is based on two publications that cover the GENEA challenge. Chapter 5 is based on one publication. First, we list which publications have been used for the chapters, including the contributions per author following CRediT (Contributor Roles Taxonomy)[1], followed by a list of publications for both journals and conferences/workshops.

Ch. 2 Wolfert, P., Robinson, N., & Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems, 52*(3), 379–389. https://doi.org/10.1109/THMS.2022.3149173

Conceptualization: P.W. and N.R.; Data curation: P.W.; Formal analysis: P.W. and N.R.; Funding acquisition: P.W. and T.B.; Investigation: P.W. and N.R.; Methodology: P.W. and N.R.; Project administration: P.W.; Supervision: N.R. and T.B.; Validation: P.W. and N.R.; Writing – original draft: P.W., N.R. and T.B.; Writing - review & editing: P.W., N.R. and T.B.;

Ch. 3 Wolfert, P., Kucherenko, T., Kjellström, H., & Belpaeme, T. (2019). Should beat gestures be learned or designed?: A benchmarking user study. *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*

Conceptualization: P.W. and T.K.; Data curation: P.W. and T.K.; Formal analysis: P.W.; Funding acquisition: H.K. and T.B.; Investigation: P.W. and T.K.; Methodology: P.W. and T.K.; Project administration: P.W.; Software: P.W. and T.K.; Supervision: H.K. and T.B.; Validation: P.W. and T.K.; Visualization: P.W. and T.K.; Writing – original draft: P.W. and T.K.; Writing - review & editing: P.W., T.K., H.K. and T.B.;

Wolfert, P., Girard, J. M., Kucherenko, T., & Belpaeme, T. (2021). To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 494–502. https://doi.org/10.1145/3462244.3479889

Conceptualization: P.W., T.K. and J.M.G.; Data curation: P.W.; Formal analysis: P.W. and J.M.G.; Funding acquisition: P.W.; Investigation: P.W., T.K. and J.M.G.; Methodology: P.W., T.K., J.M.G. and T.B.; Project administration: P.W.; Resources: P.W. and T.K.; Software: P.W.; Supervision: T.B.; Visualization: P.W. and J.M.G.; Writing – original draft: P.W., T.K., J.M.G. and T.B.; Writing - review & editing: P.W., T.K., J.M.G. and T.B.;

Ch. 4 Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., & Henter, G. E. (2021). A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 11–21. https://doi.org/10.1145/3397481.3450692

[1] https://credit.niso.org/

Conceptualization: T.K., P.J., Y.Y., P.W. and G.E.H.; Data curation: T.K., P.J. and P.W.; Formal analysis: T.K., P.J., Y.Y., P.W. and G.E.H.; Methodology: T.K., P.J., Y.Y., P.W. and G.E.H.; Project administration: T.K.; Software: T.K., P.J. and Y.Y.; Writing – original draft: T.K., P.J., Y.Y., P.W. and G.E.H.; Writing - review & editing: T.K., P.J., Y.Y., P.W. and G.E.H.;

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2022). The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 736–747. https://doi.org/10.1145/3536221.3558058

Conceptualization: T.K., Y.Y., P.W., C.V., T.N., M.T. and G.E.H.; Data curation: P.W., T.N. and M.T.; Formal analysis: T.K., P.W. and G.E.H.; Funding acquisition: T.K. and G.E.H.; Investigation: T.K., Y.Y., P.W. and G.E.H.; Methodology: T.K., Y.Y., P.W., C.V., T.N., M.T. and G.E.H.; Project administration: P.W.; Software: T.N. and M.T.; Validation: P.W. and G.E.H.; Writing – original draft: T.K., Y.Y., P.W., C.V., T.N., M.T. and G.E.H.; Writing - review & editing: T.K., Y.Y., P.W., T.N., M.T. and G.E.H.;

Ch. 5 Wolfert, P., Henter, G. E., & Belpaeme, T. (2024). Exploring the effectiveness of evaluation practices for computer-generated nonverbal behaviour. *Applied Sciences*, *14*(4). https://doi.org/10.3390/app14041460

Conceptualization: P.W., G.E.H. and T.B.; Data curation: P.W.; Formal analysis: P.W. and G.E.H.; Funding acquisition: P.W. and T.B.; Investigation: P.W.; Methodology: P.W. and G.E.H.; Project administration: P.W.; Software: P.W.; Supervision: G.E.H. and T.B.; Validation: P.W. and G.E.H.; Visualization: P.W.; Writing – original draft: P.W., G.E.H. and T.B.; Writing - review & editing: P.W., G.E.H. and T.B.;

1.5.1 List of journal publications

1. Kucherenko, T., Wolfert, P., Yoon, Y., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2024). Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022 [Just Accepted]. *ACM Trans. Graph.* https://doi.org/10.1145/3656374

2. Wolfert, P., Henter, G. E., & Belpaeme, T. (2024). Exploring the effectiveness of evaluation practices for computer-generated nonverbal behaviour. *Applied Sciences*, *14*(4). https://doi.org/10.3390/app14041460

3. Wolfert, P., Robinson, N., & Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, *52*(3), 379–389. https://doi.org/10.1109/THMS.2022.3149173

1.5.2 List of conference and workshop publications

1. Wolfert, P., De Gersem, L., Janssens, R., & Belpaeme, T. (2024). Multi-modal language learning: Explorations on learning japanese vocabulary. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 1129–1133. https://doi.org/10.1145/3610978.3640685

2. Wolfert, P., Henter, G. E., & Belpaeme, T. (2023). "am i listening?", evaluating the quality of generated data-driven listening motion. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 6–10. https://doi.org/10.1145/3610661.3617160

3. Amioka, S., Janssens, R., Wolfert, P., Ren, Q., Pinto Bernal, M. J., & Belpaeme, T. (2023). Limitations of audiovisual speech on robots for second language pronunciation learning. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 359–367. https://doi.org/10.1145/3568162.3578633

4. Neerincx, A., Leven, J., Wolfert, P., & de Graaf, M. M. (2023). The effect of simple emotional gesturing in a socially assistive robot on child's engagement at a group vaccination day. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 162–171. https://doi.org/10.1145/3568162.3576960

5. Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2022). The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 736–747. https://doi.org/10.1145/3536221.3558058

6. Wolfert, P., Kucherenko, T., Viegas, C., Yumak, Z., Yoon, Y., & Henter, G. E. (2022). Genea workshop 2022: The 3rd workshop on generation and evaluation of non-verbal behaviour for embodied agents. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 799–800. https://doi.org/10.1145/3536221.3564027

7. Janssens, R., Wolfert, P., Demeester, T., & Belpaeme, T. (2022). 'cool glasses, where did you get them?" generating visually grounded conversation starters for human-robot dialogue. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 821–825. https://doi.org/10.1109/HRI53351.2022.9889489

8. Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., & Henter, G. E. (2021). A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 11–21. https://doi.org/10.1145/3397481.3450692

9. Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Yumak, Z., & Henter, G. (2021). Genea workshop 2021: The 2nd workshop on generation and evaluation of non-verbal behaviour for embodied agents. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 872–873. https://doi.org/10.1145/3462244.3480983

10. Wolfert, P., Girard, J. M., Kucherenko, T., & Belpaeme, T. (2021). To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 494–502. https://doi.org/10.1145/3462244.3479889

11. Jonell, P., Yoon, Y., Wolfert, P., Kucherenko, T., & Henter, G. E. (2021). Hemvip: Human evaluation of multiple videos in parallel. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 707–711. https://doi.org/10.1145/3462244.3479957

12. Oetringer, D., Wolfert, P., Deschuyteneer, J., Thill, S., & Belpaeme, T. (2021). Communicative function of eye blinks of virtual avatars may not translate onto physical platforms. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 94–98. https://doi.org/10.1145/3434074.3447136

13. Wolfert, P., Deschuyteneer, J., Oetringer, D., Robinson, N., & Belpaeme, T. (2020). Security risks of social robots used to persuade and manipulate: A proof of concept study. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 523–525. https://doi.org/10.1145/3371382.3378341

14. Wolfert, P., Kucherenko, T., Kjellström, H., & Belpaeme, T. (2019). Should beat gestures be learned or designed?: A benchmarking user study. *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*

Background and Related Work

2

Background and Related Work

In order to understand the next three chapters on the evaluation and generation of nonverbal behaviour for ECAs, we first explore the background and related work of the field. In this chapter we look at the different approaches that have been used over the years to learn, synthesise and evaluate nonverbal behaviour for ECAs. We deep dive into history by looking at the first approaches for nonverbal behaviour generation using rule based generation. Then, we look at data driven approaches that have gained popularity over the recent years, and at the most used datasets.

2.1 Nonverbal Behaviour

Human communication involves a large nonverbal component, with some suggesting that a large portion of communicative semantics is drawn from non-linguistic elements of face-to-face interaction (Knapp et al., 2013). Nonverbal behavior can be broken down into several elements, such as posture (think of our pose when we experience negative emotions), gestures (i.e. pointing at an object of interest), facial expressions (i.e. expressing an emotion), gaze (i.e. indicating our attention), proxemics (i.e. how comfortable are we with our interlocutor?), and haptics (i.e., touch during communicative interactions). All these elements convey different types of meaning, which can complement or alter the semantic component of communication. Even minimal elements can provide a marked contribution to the interaction. For example, eye blinking with head nodding has been found to influence the duration of a response in a Q&A session between human subjects and a robot (Hömke et al., 2018) Humans can intentionally display nonverbal behaviour, for example through gesturing or by using facial expressions. This type of behaviour could be classified as nonverbal communication, however, as pointed at by Hall et al. (2019), the terms nonverbal communication and nonverbal behaviour are used interchangeably most of the time. Besides adding meaning to a conversation, nonverbal behaviour can tell us a lot about the individual we are interacting with, one example being that we unconsciously send signals about our identity and intentions (Hall et al., 2019).

2.2 Gestures

A significant component involved in nonverbal communication is the use of gestures –movements of the hands, arms, or body– to emphasize a message, communicate an idea, or express a sentiment (Knapp et al., 2013). Humans often use gestures in daily life, such as to point at objects in our visual space, or to signal the size of an object. Before we dive into the different dimensions of gestures as defined by McNeill, we first take a look at what a constitutes a gesture. Kendon (1980) describes how gesticulation unfolds over time. The moment we start moving a forelimb to make a gesture motion until moving it back into a rest position, is described as a *gesticular unit*. During a gesticular unit, the limb in question can perform one more more phrases of gesticulation. A *G-Phrase* or phrase of gesticulation, shows, what Kendon calls, a distinct peaking of 'effort'. This is defined as the *stroke*. A stroke is preceded by a *preparation* phase and followed by a *recovery* or *return phase*. Following this, one can also say that a gesture is triphasic. So, one gesture unit can contain multiple gesture phrases, where each phrase has a start and an end in the resting position of the limb(s) in case.

McNeill (1992), known for their work on gesture and speech, writes the following definition: "All visible movements by the speaker are first differentiated into gestures and non-gestures; the latter comprise self-touching (e.g., stroking the hair) and object manipulation. The rest are considered gestures are classified as to type." Following that, they define two types of gestures. The ones that are imagistic, and non-imagistic. Gestures either depict imagery or not. Additionally, McNeill states that imagistic gestures are often triphasic, whereas non-imagistic gestures such as beat gestures, are not. Following these definitions, four kinds of co-speech gestures are identified: iconic gestures, metaphorical gestures, beat gestures, and deictic gestures. Iconic and metaphorical gestures both carry meaning and are used to visually enrich our communication (Kendon, 1980). An iconic gesture can be an up and down movement to indicate, for example, the action of slicing a tomato. Instead, a metaphoric gesture can involve an empty palm hand that is used to symbolize 'presenting a problem'. In other words, metaphoric gestures have an arbitrary relation to the concept they communicate, and iconic gestures have a form that is visually related to the concept being communicated. Iconic and metaphoric gestures not only differ in terms of content and presentation, but are also processed differently in the brain (Straube et al., 2011). Beat gestures are gestures we use when we aim to emphasize a message. The last kind are deictic gestures: deictic gestures are used to point out elements of interest or to communicate directions.

It is important to mention that these four kinds can not be seen as individual categories, but more as dimensions. McNeill (2019) argues for a dimensional view on gestures, and provides the following dimensions: "iconicity, metaphoricity, deixis, temporal highlighting (beats), and social interactivity." We see that the earlier defined 'categories' appear as dimensions, but using

these dimensions, a gesture can be characterized alongside multiple dimensions. For example, the size of an object can be emphasized by also including a beat at the onset of the gesture. However, many in the gesture generation scene still describe the categories mentioned earlier, and seem to sometimes ignore the dimensionality of gestures.

There is a plethora of research covering the use of co-speech gestures in human communication, but much about the interaction between speech and gestures stays unknown (Wagner et al., 2014). We know that gesture and speech intertwined, and affect each other (Kelly et al., 2010). For example, de Ruiter et al. (2012) discusses and studies the tradeoff hypothesis. The tradeoff hypothesis states that when gesturing is impaired, we start to rely more on speech, and when our speech is impaired, we start to rely more on gesturing. Although they do not find evidence for the second part of the hypothesis in their study, they find that the gesture rate for beat gestures is not affected throughout their experiments. This could be seen as evidence that not all gesturing has pure communicative function. Others find that beat gestures facilitate speech and word recall (Igualada et al., 2017; Lucero et al., 2014) and are the most frequent type of gesture (Chui, 2005; Kong et al., 2015; McNeill, 1992). When and how we gesture is also dependent on the context of the situation. Holler and Stevens (2007) shows that referential gesturing is dependent on the existing common ground, i.e., when there was no common ground about object' sizes, speakers would gesture, whereas they would rely on verbal information only when this information was already part of the common ground. Additionally, pointing gestures have been found to facilitate learning (Lucca & Wilbourn, 2018). Through gesture features, we can for example also deduce the affective state, such as pleasure and arousal (Kipp & Martin, 2009).

2.3 Gesturing in Embodied Conversational Agents

As nonverbal behavior plays an important role in human-human interaction, researchers put substantial efforts into the generation of nonverbal behavior for ECAs. ECAs, such as social robots, can display a range of nonverbal behaviors, including the ability to make gesture-like movements (Bartneck et al., 2020; Breazeal et al., 2005; Saunderson & Nejat, 2019). The use of co-speech gestures in communication with humans by ECAs can influence the perception and understanding of the conveyed message (Allmendinger, 2010; Bremner, Pipe, et al., 2009). For example, participants recalled more facts from a narrative told by an ECA, when the ECA made use of deictic and beat gestures compared to when the ECA did not make use of gesticulation (Huang & Mutlu, 2013, 2014). As another example, humans are more willing to cooperate when an ECA showed appropriate gesturing (consisting of deictic, iconic, and metaphoric gestures) in comparison to when an ECA did not use gestures or when the gestures did not match the verbal utterances (Salem, Eyssel, et al., 2013). Gestures are particularly salient in humanoid robotics,

i.e., when the ECA is physically embodied. Robots can be perceived to be more persuasive when they combine gestures with other interactive social behaviors, such as eye gaze, in comparison with when they do not use either of these techniques (Chidambaram et al., 2012; Ghazali et al., 2019; Ghazali et al., 2018; Ham et al., 2015). This demonstrates the impact nonverbal behavior from ECAs can have on people and its importance for consideration in human-agent interactions.

2.4 Gesture Generation

We can roughly identify two approaches to generating co-speech gestures. One approach is to take into account communicative intent, and generate behaviour given this communicative intent. The other approach is to learn or derive a mapping, going from one modality to the other, e.g. learning a mapping based on speech-audio to generate co-speech gestures.

Following their earlier work on a rule-based system that was able to generate non-verbal behaviours tied to one specific type of interaction, (Cassell et al., 2000) proposed a model that was able to generate behaviour that takes into account communicative goals. This model was also rule-based. Another way to generate meaningful gestures that would take account of communicative intent and goals, is parsing input text, and plan gestures given the 'understanding' of the text (Kopp et al., 2004). Data-based systems were introduced to generate more dynamic looking behaviour. For example, (Bergmann & Kopp, 2009; Bergmann et al., 2010) introduced GNetIc, a Bayesian decision network fed with annotated data to generate iconic gestures.

2.4.1 Rule Based Generation

One of the earliest examples of rule-based gesture generation is the work by (Cassell et al., 1994). Rules are used for the mapping from speech-text to gestures and facial expressions. A drawback of a rule-based approach is that it does not scale well, and requires a lot of manual feature engineering. Later, these separate systems found their way into the BEAT toolkit, which featured a rule-based system that relies on linguistic and contextual analysis of the input text to suggest behaviours (Cassell et al., 2001). Marsella et al. (2013) proposed a rule-based system that relied on acoustic features in combination with lexical content. In 2006, Behaviour markup language (BML) was introduced, which is an adaptation of the Extensible markup language (XML) format to describe all kinds of behaviours (verbal and non-verbal) (Kopp et al., 2006). The aim was to unify the approach of how behaviour in ECA is established. A typical piece of BML script describes the occurrence of a behaviour, the timing of such behaviour and the form of the behaviour. BML stands between behaviour generation engines, that define which behaviour

needs to be performed and behaviour realisers, that can read BML to directly drive ECAs.

Another mapping approach, that does not require learning, is using a statistical model. Neff et al. (2008) introduced an algorithm that considers several annotated gesture features, to integrate gestures with speech. (Levine et al., 2009, 2010) introduced gesture controllers that rely on reinforcement learning and acoustic features. Although a learning component is included in this work, it is only used to probabilistically select a motion sequence from a library. Hence, the motion sequences are not generated, but selected from this library.

2.4.2 Data Driven Generation

Instead of relying on hand-coding rules, gesture generation systems can also be created from human conversational data, this is known as the data-driven approach (Bergmann & Kopp, 2009; Levine et al., 2009). These data-driven methods have predominantly relied on neural networks for synthesizing gestures. Paired with the rise of deep learning techniques, data-driven methods are capable of unprecedented generalization, an invaluable property when generating high dimensional temporal output. Data-driven approaches using neural networks are capable of generating more dynamic and unique gestures, but this does heavily depend on the available training data and the type of neural networks that are used. Some approaches learn a mapping from acoustic features of speech signals to gesture (Hasegawa et al., 2018; Kucherenko et al., 2019). Audio signal-based methods are now much better at creating dynamic and fluent beat gestures, whereas text-based methods show an improved generation of iconic and metaphoric gestures. However, relying on only acoustic features of the speech audio means that semantic details are lost, hence these approaches often only generate beat gestures. Work by Kucherenko et al. (2020) combines neural networks for beat gesture generation with sequential neural networks for generating iconic gestures, dispensing with the need for a rule-based hybrid approach. Yoon et al. (2019), trained an encoder-decoder neural network on combinations of subtitles and human poses extracted from public TED(x) videos. This allowed the network to learn a relationship between written language, extracted from the video's subtitles, and gesture and was used to generate beat and iconic gestures for a humanoid robot. However, an in-depth evaluation of the different categories of gestures generated by the system was not part of the study. This method was a notable advance in gesture generation, given that videos contain a wealth of human conversational data and are abundantly available. More recently, researchers started picking up diffusion models for gesture synthesis (Alexanderson et al., 2023; Ao et al., 2023; Mehta et al., 2023). A more in depth review on the field of gesture generation, especially considering deep learning, can be found in (Nyatsanga et al., 2023). The data used to build data-driven gesture generation can vary, where some use data collected from many individuals (Yoon et al., 2019), others make use of data sets containing a single actor (Ferstl

et al., 2019). The latter is often used in the context of virtual reality, where the data is collected from a single actor, and the data is used to animate a virtual avatar.

2.4.3 Datasets

There are several datasets that have been used for learning co-speech gesturing, and here we discuss the ones that are used in this thesis. Work by Nyatsanga et al. (2023) provides an overview of the most used datasets in the field of data-driven co-speech gesture generation.

Trinity Speech-Gesture

The Trinity Speech-Gesture dataset was released in 2018 (Ferstl & McDonnell, 2018). This dataset features a single male actor, who is a native speaker of English. In the dataset, the actor speaks freely and spontaneous, while also displaying gesture motion. There are a total of 6 (244 minutes) hours of recordings, split up in 23 takes. Recordings were made in 3D using 20 Vicon cameras and 53 body markers, at 59.94 Frames per second (fps). The final output format were joint rotations. Included with the gesture motion was the speech audio.

For the GENEA 2020 Challenge, the audio was transcribed and the motion data was cleaned (Kucherenko, Jonell, Yoon, Wolfert, & Henter, 2021).

Talking with Hands

The Talking with Hands dataset was released in 2021 and originally consists of 32 sessions. The subset that contains both motion and audio covers 17 sessions. In contrast to the Trinity Speech-Gesture dataset, this dataset contains dyadic conversations, featuring multiple subjects. Each session is approximately 10 minutes long. The original dataset has a length of 50 hours, but not all takes have been released. Motion is recorded in 3D joint angle rotations using 24 cameras in 90 Fps. For this, 24 OptiTrack Prima 17W cameras were placed in an area measuring 3 by 3 meters. This dataset is one of the few that also feature hiqh quality finger data. The dataset is recorded

For the GENEA 2022 Challenge, the audio was described, and parts of the dataset were held out due to motion error or bad audio.

2.5 Listening Behaviour Generation

Since we are concerned with listening behaviour generation in the last chapter of this thesis, we provide background on listening behaviour generation. Listening is an essential aspect of human-agent interaction, and studies have shown that virtual agents who pretend to listen can enhance engagement during an interaction (Heylen et al., 2011). For instance, Buschmeier and Kopp (2018) showed that when humans interacted with an attentive agent, they were more likely to provide listener feedback and rated the agent as more helpful. Maatman et al. (2005) proposed a model that generates listening behaviour based on available features during a conversation. Their system extracts audio and body posture features to drive the listening behaviour. Another approach by Gillies et al. (2008) utilised input audio from the speaker to generate listening behaviour through motion graphs, where existing motion clips are combined to match new audio input. Mlakar et al. (2014) introduced a framework and scripting method to synthesise both verbal and nonverbal motion, that entails both gestures and listening. Poppe et al. (2010) developed rule-based strategies for generating listening behaviour based on the speaker's speech and gaze, including vocal back channelling. A similar approach in terms of selecting new listening behaviours and sequences can be found in (Gómez Jáuregui et al., 2021). They used a multi-modal corpus of interviews to generate listening behaviour in a virtual agent conducting interviews. Participants perceived the interviewer as affiliative when the interviewer would mirror their posture. An example of generating listening head behaviour is the work by Jonell, Kucherenko, Henter, and Beskow (2020). They generated interlocutor-aware facial gestures using nonverbal and verbal input from both the interlocutor and agent, using a generative approach.

2.6 Objective Evaluations

A central component for any method that can generate human-like behavior is the ability to evaluate the quality of the generated signals. To date, researchers make use of a variety of different methods to evaluate gesture generation systems. One way is to use objective evaluations, often consisting of metrics for the joint speed, joint trajectories, jerk, or the Frechet Gesture Distance. The objective metrics that are often reported are not necessarily the same metrics that are used to train neural networks.

Loss functions used for training networks only tell how close the generated stimuli are to the ground truth, and they do not provide information on whether the generated motion is dynamic or natural enough. We will now take a look at the objective metrics that are reported on throughout this thesis.

Average acceleration and jerk

The third time derivative of the joint positions is called *jerk* and can be formulated mathematically as $\text{jerk}(\boldsymbol{x}) = \boldsymbol{x}'''(t)$. The average value of the absolute magnitude of the jerk is commonly used to quantify motion smoothness (Kucherenko et al., 2019; Morasso, 1981; Uno et al., 1989). We report average values of absolute jerk (defined using finite differences) averaged across all test motion segments. A perfectly natural system should have average jerk very similar to natural motion.

We also evaluated the same measure, but computed using the absolute value of the acceleration $\text{acc.}(x) = x''(t)$ instead of the jerk. Again, we expect natural-looking motion to have similar average acceleration as in the reference data.

Comparing speed histograms

The distance between speed histograms has also been used to evaluate gesture quality (Kucherenko et al., 2019, 2020), since well-trained models should produce motion with similar properties to that of the actor it was trained on. In particular, it should have a similar motion-speed profile for any given joint. This metric is based on the assumption that synthesised motion should follow a similar speed distribution as the ground truth motion. We can evaluate this similarly by calculating speed-distribution histograms for all conditions we want to compare to the speed-distribution of natural motion, by computing the Hellinger distance (Nikulin, 2001).

$$H(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}) = \sqrt{1 - \sum_i \sqrt{h_i^{(1)} \cdot h_i^{(2)}}}, \tag{2.1}$$

between the histograms $\boldsymbol{h}^{(1)}$ and $\boldsymbol{h}^{(2)}$. Lower distance is better.

Canonical correlation analysis

Canonical correlation analysis (CCA) (Thompson, 1984) is a form of linear subspace analysis, and involves the projection of two sets of vectors onto a joint subspace. CCA has been used to evaluate gesture-generation models in previous work (Bozkurt et al., 2015; Lu et al., 2020; Sadoughi & Busso, 2019).

The goal of CCA is to find a sequence of linear transformations of each variable set, such that the Pearson correlation between the transformed variables is maximised. This correlation is what we d as a similarity measure, and it is reported as global CCA in the results section. A high value is considered better.

Fréchet gesture distance

Recent work by Yoon et al. (2020) proposed the Fréchet gesture distance (FGD) to quantify the quality of generated gestures. This metric is based on the FID metric used in image-generation studies to quantify the realism and diversity of images generated by generative adversarial networks (GANs) (Heusel et al., 2017) and can be written

$$\mathrm{FGD}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = ||\boldsymbol{\mu}_r - \boldsymbol{\mu}_g||^2 + \mathrm{tr}(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2}). \tag{2.2}$$

Here, $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$ are the first and second moments of the latent-feature distribution $\boldsymbol{Z}_r$ of the human motion-capture data $\boldsymbol{X}$, whereas $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the first and second moments of the latent-feature distribution $\boldsymbol{Z}_g$ of the generated gestures $\hat{\boldsymbol{X}}$. $\boldsymbol{Z}_r$ and $\boldsymbol{Z}_g$ were extracted by the same feature extractor, which was obtained as the encoder part of a motion-reconstructing autoencoder. Lower values are better.

System ranking comparison

A good objective metric might help in evaluating the performance of a system, especially when such a metric correlates with a subjective measure. To get more insight into whether the objective metrics in our study may be used as a proxy for subjective evaluation results, we calculated the correlation between the ranking of the conditions on median human-likeness, and the result on the objective metrics listed above. For this, we used Kendall's τ rank correlation coefficient, and associated statistical tests (Kendall, 1970).

Of the objective metrics we took into account for both challenges, only CCA compares output poses directly to the corresponding reference motion-capture poses. All other metrics are invariant to permutation, in the sense that changing the order of the different sequences (mismatching them with other speech/reference motion) will not change the value. They thus cannot measure appropriateness, which is why we only consider how those metrics correlate with human-likeness scores.

2.7 Subjective Evaluations

In addition to using objective evaluations, subjective evaluations are a key part of evaluating generated nonverbal behaviour. For this, one needs to setup a user study, where human participants evaluate the performance of the gestures used by an ECA. Examples of dimensions on which the performance of an ECA is evaluated, are, for example, the perceived naturalness of the generated motion, the perceived appropriateness of the gestures' timing, 'speech-gesture correlation' or 'naturalness' (Ishi et al., 2018; Levine et al., 2009). These are often evaluated using several items in one Likert Scale. In

human-robot interaction (Yoon et al., 2019), researchers have used questionnaires for general robot evaluation, such as the Godspeed questionnaire, or a selection of items from such instruments. The Godspeed questionnaire can evaluate the perception of ECAs in a non-domain-specific measurement, and quantifies the human likeness, animacy, likability, and perceived intelligence of ECAs (Bartneck et al., 2009). Other methods measure the effect that the gesticulation of an ECA has on the user, such as listener's comprehension and recall of spoken material (Huang & Mutlu, 2013, 2014).

2.7.1 Analysis of Subjective Evaluations used in Gesture Generation

To gain a deeper understanding on the usage of subjective evaluation methods, we reviewed work on gesture generation for ECAs. We followed the PRISMA protocol (Moher et al., 2010) to identify and assess evaluation methods used in co-speech gestures. Central to this section are three research questions:

1. What methods are used to evaluate co-speech gesture generation?
2. Which methods can be considered the most effective for assessing co-speech gestures?
3. What methods and related metrics should be adapted to create a standardized evaluation or reporting protocol?

2.7.2 Methods

Search Strategy

We consulted three databases for data extraction: IEEE Explore, Web of Science, and Google Scholar. IEEE Explore was selected given that it captures a substantial number of publications in computer science and engineering. Web of Science and Google Scholar were used because they provide access to multiple databases with a wide coverage extending beyond computer science and engineering. We conducted independent data extraction steps to reduce the chance of relevant papers being missed from the review, which included inter-rater checks on the included records. The databases were queried using four different keyword combinations, where the search engine would add 'AND' between keywords: 1) "gesture generation for social robots", 2) "co speech gesture generation", 3) "non verbal gesture generation", and 4) "nonverbal behavior generation".

Eligibility – Inclusion and Exclusion

We set up inclusion criteria to determine which work should be part of our analyses:

1. The ECA paper must report on gesture generation on either a robot or an embodied agent.

2. The ECA system must be humanoid in nature, with one or two human-like arms and/or hands that can be used to gesture information or messages to the human.

3. The ECA system must display multiple gestures (i.e., a minimum of 2 different gestures, one of which must be a beat, iconic, metaphoric or deictic gesture).

4. Gestures created by the ECA system must be those that would be seen during a multi-modal social interaction.

5. The ECA paper must report on a user study (i.e., not evaluated using technical collaborators or authors) in a laboratory, in the wild, or performed remotely through online platforms.

6. The ECA system must be evaluated by a human rater on its performance (either directly or indirectly).

To narrow down our search results, we used the following exclusion criteria:

1. The paper contains a non-humanoid agent that lacks a typical human-like hand for making a gesture.

2. The paper does not have a clear focus on evaluation of co-speech gestures, i.e., secondary measures that is less than 50% of the paper.

3. The paper only covers beat gesture generation.

4. The paper is either unpublished, a doctoral dissertation, a review, a technical paper or pre-print.

5. The paper is not written in English.

Extracted records that only included beat gesture generation were recorded but excluded from our main analysis, as these records rely on audio inputs for the generation of beat gestures. Hence, these beat gesture generation systems do not take semantic information into account. Instead, a separate analysis outside the PRISMA protocol is provided to consider work on beat gestures only, as we do consider the work on beat gesture generation important.

2.7.3 Results

In this section, we discuss the results of our literature search. First, we discuss the found articles, followed by a discussion on the usage of different ECAs. Then, we discuss the characteristics of participant samples in experiments, the design of the experiments, and the use of objective and subjective evaluations. At the end, we present the results of our analysis of papers that only incorporated beat gesture generation.

Selected Articles

The initial search conducted across three separate databases resulted in 295 papers, which contained 92 duplicate records. A total of 203 papers were screened for their titles and abstracts for an initial exclusion step, resulting in 113 papers being omitted for not meeting all the inclusion criteria. The 90 remaining papers were assessed in detail by reviewing the main text for eligibility. The 68 non-eligible papers met one or more exclusion criteria, and were therefore discarded. This resulted in 22 papers that met all inclusion criteria and none of the exclusion criteria. Figure 2 shows the PRISMA flow chart with the results of this process. Extracted information from the manuscripts included publication year, venue, design and conditions, method of generation, objective metrics, subjective metrics, type of ECA, evaluation type (online, in the wild, or in a laboratory), participants, characteristics of participants, and other important notes related to the experiment.

Embodied Conversational Agents

In the 22 included studies, 16 studies (73%) used different human-like robots, such as NAO (n = 3, 14%), ASIMO (n = 3, 14%) or Wakamaru (n = 2, 9%). Only 6 (27%) reported the use of a virtual agent (*viz.* (Ishii et al., 2018; Levine et al., 2010; Mlakar et al., 2013; Neff et al., 2008; Rojc et al., 2017; Xu et al., 2014)). All the virtual agents were modelled in 3D as a virtual human, and there were no consistent features across the agents between studies. Of the 6 studies, 4 used female avatars (Ishii et al., 2018; Mlakar et al., 2013; Rojc et al., 2017; Xu et al., 2014), 1 used a male avatar (Neff et al., 2008) and 1 study used both(Levine et al., 2010). Half of the studies that used avatars, showed only the upper body (Mlakar et al., 2013; Neff et al., 2008; Rojc et al., 2017), whereas the other half showed full-body avatars (Ishi et al., 2018; Levine et al., 2010; Xu et al., 2014). Specific descriptions of the hands were not provided in all the studies that used avatars. In 19 (87%) studies, the ECA performed iconic gestures, combined with other gestures (Aly & Tapus, 2013; Bennewitz et al., 2007; Huang & Mutlu, 2013, 2014; Ishi et al., 2018; Ishii et al., 2018; Q. Le et al., 2012; Q. A. Le & Pelachaud, 2012; Mlakar et al., 2013; Ng-Thow-Hing et al., 2010; Pérez-Mayos et al., 2020; Rojc et al., 2017; Salem, Eyssel, et al., 2013; Salem, Kopp, & Joublin, 2013; Salem et al., 2011, 2012; Shimazu et al., 2018; Yoon et al., 2019).

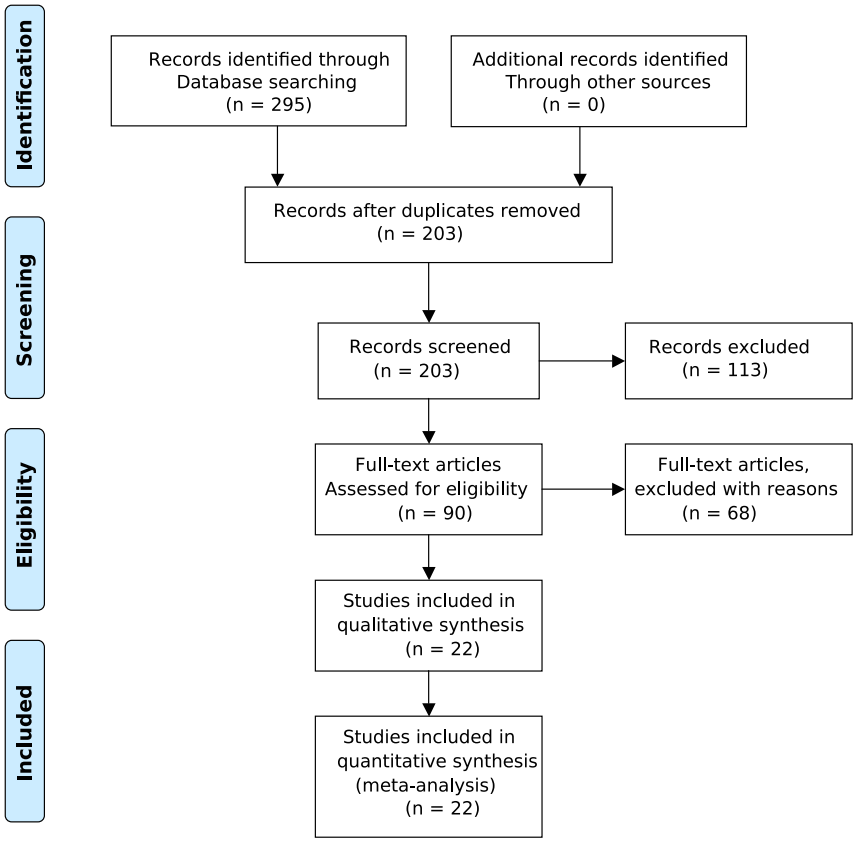


Figure 2.1 PRISMA Flow Chart

Metaphoric gestures, with other gestures, are used in 17 (77%) studies (Aly & Tapus, 2013; Huang & Mutlu, 2013, 2014; Ishi et al., 2018; Ishii et al., 2018; Q. Le et al., 2012; Q. A. Le & Pelachaud, 2012; Mlakar et al., 2013; Neff et al., 2008; Ng-Thow-Hing et al., 2010; Rojc et al., 2017; Salem, Eyssel, et al., 2013; Salem, Kopp, & Joublin, 2013; Salem et al., 2011, 2012; Xu et al., 2014; Yoon et al., 2019). Deictic gestures, with other gesture types, play a key role in 13 (59%) of the reviewed studies (Bennewitz et al., 2007; Huang & Mutlu, 2013, 2014; Ishi et al., 2018; Ishii et al., 2018; H. Kim et al., 2012; Q. Le et al., 2012; Q. A. Le & Pelachaud, 2012; Mlakar et al., 2013; Ng-Thow-Hing et al., 2010; Rojc et al., 2017; Salem, Eyssel, et al., 2013; Salem, Kopp, & Joublin, 2013; Salem et al., 2011, 2012; Yoon et al., 2019). Lastly, 17 (77%) studies included iconic, metaphoric *and* beat gestures (Aly & Tapus, 2013; Bennewitz et al., 2007; Huang & Mutlu, 2013, 2014; Ishi et al., 2018; Ishii et al., 2018; H. Kim et al., 2012; Q. Le et al., 2012; Q. A. Le & Pelachaud, 2012; Levine et al., 2010; Neff et al., 2008; Ng-Thow-Hing et al., 2010; Pérez-Mayos et al., 2020; Rojc et al., 2017; Shimazu et al., 2018; Xu et al., 2014; Yoon et al., 2019). Half of the studies had the ECA perform 'random gestures' that were included in the evaluation (i.e., gestures that had no alignment between gestures and speech). Other studies (n = 4) had the ECA present the user with a variety of different nonverbal behavior schemes, such as gestures that were based on text, speech, or a combination of the two(Ishii et al., 2018; Pérez-Mayos et al., 2020; Salem, Eyssel, et al., 2013; Shimazu et al., 2018).

Participants

The number of participants per study ranged from 13 to 250 in total (mean = 50, SD = 50, median = 35). In these papers, 19 (86%) were conducted in the laboratory, and 3 (14%) were conducted either online through Amazon mechanical turk (AMT) (n = 2) and 1 during an exhibition (i.e., 'in the wild'). For the 12 (54%) studies that did report the mean age of the participants, the mean reported age across all studies was 30.10 years of age (Standard deviation (SD) = 6.6). The remaining 11 (46%) did not provide demographic data for gender and age. Relating to trial location, 16 (73%) of studies were performed outside English-speaking countries, with the top 3 countries being Germany (n = 5), Japan (n = 3), and France (n = 3). For participant recruitment, 6 (27%) of the studies reported the use of university students –a so-called *convenience sample*– to evaluate gesture generation. Table I provide a more detailed overview of the different studies, countries of origin, and characteristics.

Research Experiment and Assessment

In research design, 16 (68%) of the studies used a within-subject design and 7 (32%) used a between-subject design. Most (n = 18, 82%) studies invited participants to a university research laboratory to have an interaction with an ECA. Other methods used AMT (n = 2, 9%). With use in 9 (41%) studies,

Table 2.1 Participants in Studies

Study	Country	Gender	Mean Age (SD)	N	Characteristics	Lab/Remote Evaluation
Yoon et al. (2019)	South Korea	23M/23F	37 (-)	46	45 USA, 1 Australia	AMT
Pérez-Mayos et al. (2020)	Spain	-	-	50	Non-native English Speakers	In Lab
Ishii et al. (2018)	Japan	-	-	10	Age + Gender not specified	In Lab
Ishi et al. (2018)	Japan	-	-	20	Age + Gender not specified	In Lab
Shimazu et al. (2018)	Japan	-	-	13	-	In Lab
Rojc et al. (2017)	Slovenia	22M/8F	-	30	-	In Lab
Xu et al. (2014)	U.S.A.	-	-	250	One 'worker' per comparison	AMT
Huang and Mutlu (2014)	U.S.A.	16M/13F	22.62 (4.35)	29	Convenience Sample	In Lab
Salem, Kopp, and Joublin (2013)	Germany	10M/10F	28.5 (4.53)	20	Native German Speakers	In Lab
Aly and Tapus (2013)	France	14M/7F	21-30	21	Convenience Sample	In Lab
Mlakar et al. (2013)	Slovenia	23M/7F	26.73 (4.88)	30	Convenience Sample	In Lab
Huang and Mutlu (2013)	U.S.A.	16M/16F	24.34 (8.64)	32	Convenience Sample	In Lab
Salem, Eyssel, et al. (2013)	Germany	30M/32F	30.90 (9.82)	62	Convenience Sample	In Lab
Salem et al. (2012)	Germany	30M/30F	31 (10.21)	60	Native German Speakers	In Lab
H. Kim et al. (2012)	South Korea	-	-	65		In Lab
Q. A. Le and Pelachaud (2012)	France	36M/27F	37 (12.14)	63	Convenience Sample	In Lab
Q. Le et al. (2012)	France	-	-	63	French Speakers	In Lab
Salem et al. (2011)	Germany	20M/20F - 20M/20F	31.31 (10.55)/31.54(10.96)	81	Two Studies	In Lab
Levine et al. (2010)	U.S.A.	21M/14F	23 (-)	35	Convenience Sample	In Lab
Ng-Thow-Hing et al. (2010)	U.S.A.	-	-	54	-	In Lab
Neff et al. (2008)	U.S.A.	20M/6F	24-26 (-)	26	Non-experts	In Lab
Bennewitz et al. (2007)	Germany	-	-	-	-	Exhibition

'naturalness' was the most common metric for evaluation in generated gestures. This was followed by synchronization (n = 6, 27%), likability (n = 4, 18%), and human-likeness (n = 2, 9%). 2 studies (9%) (Salem, Kopp, & Joublin, 2013; Xu et al., 2014) asked participants to choose which audio track matched best with a given generated gesture sequence. 9 (41%) studies made use of models that learn to generate co-speech gestures. When assessing generated gestures, 16 (73%) studies used questionnaires as a tool to evaluate ECA gesture performance. Only 1 study (Salem, Kopp, & Joublin, 2013) included a previous iteration of their gesture model for evaluation. 4 studies (18%) used a ground truth as part of the gesture generation evaluation. 3 studies (13%) relied on pairwise comparisons, such as two or more videos put side by side with the user selecting the video that best matches with the speech audio, e.g., (Levine et al., 2010; Ng-Thow-Hing et al., 2010; Pérez-Mayos et al., 2020). Other evaluation methods involved robot performance, e.g., (Huang & Mutlu, 2013, 2014).

Objective and Subjective Evaluation

Table 2.2 provides a summary of studies that involved objective evaluation. It also includes the type of agents that were used, as well as the number of speakers in a dataset (when applicable) and the setting of the speakers in the conversation. Only 5 studies (23%) involved some form of objective evaluation metrics as a key method in their evaluation. Other metrics included variations on the Mean squared error (MSE)) (n = 1, 4.5%) between the generated and ground truth gestures, and qualitative analyses of joint velocities and positions (n = 2, 9%). In total, 10 (45%) studies used a data-driven generation method, but only 3 studies (14%) reported outcomes of their objective metrics used for tuning their models. Only 3 (14%) studies reported the results of their objective metrics relating to their model performance. 7 studies (32%) relied on data featuring single speakers. In addition to that, 7 studies (32%) relied on data showing 2 or more speakers. The remainder did not report on the setting of the data or the number of speakers in their dataset.

Table 2.3 provides a detailed overview of study design, conditions, and subjective evaluation methods. Fewer studies used between-group design (n = 6, 27%) compared to within-group design (n = 16, 73%). Most were evaluated using questionnaires (n = 16, 73%) followed by pairwise comparisons (n = 3, 14%) and other methods (n = 4, 18%) such as preference matching (matching audio with video) and recalling facts from a story told by the agent.

Study	Generation Method	Objective Metrics	Agent	#N Speakers	Setting
Yoon et al. (2019)	Data Driven	Variation on Mean Squared Error	NAO	1295	Single
Pérez-Mayos et al. (2020)	Rule Based	-	REEM-C	2	Single
Ishii et al. (2018)	Data Driven	-	Virtual Agent (3D)	24	Conversation (two)
Ishi et al. (2018)	Data Driven	-	Android Erica	8	Conversation (three)
Shimazu et al. (2018)	Data Driven	Log-likelihood of generated motion	Pepper	119	Single
Rojc et al. (2017)	Hybrid	-	Virtual Agent (3D)	5	Multiple
Xu et al. (2014)	Rule Based	-	Virtual Agent (3D)	5	Single
Huang and Mutlu (2014)	Data Driven	-	Wakamaru	16	Conversation (two)
Salem, Kopp, and Joublin (2013)	Rule Based	Qualitative Analysis of Joint Positions	ASIMO	-	-
Aly and Tapus (2013)	Rule Based	-	NAO	-	-
Mlakar et al. (2013)	Data Driven	-	Virtual Agent (3D)	4	Multiple
Huang and Mutlu (2013)	Rule Based	-	Wakamaru	8	Conversation (two)
Salem, Eyssel, et al. (2013)	Rule Based	-	ASIMO	-	-
Salem et al. (2012)	Rule Based	Qualitative Analysis of Joint Positions	ASIMO	-	-
H. Kim et al. (2012)	Rule Based	-	Industrial Service Robot	1	Single
Q. A. Le and Pelachaud (2012)	Rule Based	-	NAO	-	-
Q. Le et al. (2012)	Rule Based	-	NAO	-	-
Salem et al. (2011)	Rule Based	-	ASIMO	-	-
Levine et al. (2010)	Data Driven	Cost Function on Kinematic Parameters	Virtual Agent (3D)	1	Conversation (two)
Ng-Thow-Hing et al. (2010)	Rule Based	-	ASIMO	4	Single
Neff et al. (2008)	Data Driven	-	Virtual Agent (3D)	2	Single
Bennewitz et al. (2007)	Rule Based	-	Fritz	-	-

Table 2.2 Objective Evaluation Methods

Study	Design	Conditions	Gesture Types	Evaluation	Questionnaire items
Yoon et al. (2019)	Within-subject	Ground truth, proposed method, nearest neighbors, random or manual	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Anthropomorphism, Likability, Speech-gesture correlation
Pérez-Mayos et al. (2020)	Within-subject	Part-of-Speech-Based, Prosody-Based, Combined	Iconic, Beat	Pairwise + Questionnaire	Timing, Appropriateness, Naturalness
Ishii et al. (2018)	Within-subject	None, Random, Proposed Method	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Naturalness of Movement, Consistency in utterance and movement, likability, humanness
Ishi et al. (2018)	Within-subject	No hand motion, Direct Human mapping, Text-based gestures, Text-based + prosody-based gestures	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Human-likeness, Gesture-speech suitability, Gesture-Naturalness, Gesture-Frequency, Gesture-timing
Shimazu et al. (2018)	Within-subject	Ground truth, seq2seq, seq2seq(model) + semantic, seq2seq_tts + semantic	Iconic, Beat	Questionnaire	Naturalness, Skill of presentation, Utilization of gesture, Vividness, Enthusiasm
Rojc et al. (2017)	Within-subject	Text+Speech (no avatar), Gestures	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Content Match, Synchronization, Fluidity, Dynamics, Density, Understanding, Vividness
Xu et al. (2014)	Within-subject	Hands never go into relax position, hands always go into rest position	Beat, Metaphoric	Match preference	N.A.
Huang and Mutlu (2014)	Between-subject	Learning-based, unimodal, random, conventional	Iconic, Beat, Deictic, Metaphoric	Questionnaire + Retelling Performance	Immediacy, Naturalness, Effectiveness, Likability, Credibility
Salem, Kopp, and Joublin (2013)	Within-subject	Old version, new version of model	Iconic, Deictic, Metaphoric	Match preference	N.A.
Aly and Tapus (2013)	Within-subject	Introverted versus Extraverted Robot, Adapted Speech and Behavior versus Adapted Speech	Iconic, Beat, Metaphoric	Questionnaire	24 questions on personality, interaction with the robot, speech, and gesture synchronization and matching
Mlakar et al. (2013)	Between-subject	Virtual avatar versus iCub robot	Iconic, Deictic, Metaphoric	Questionnaire	Content Matching, Synchronization, Fluidness, Speech-Gesture Matching, Execution Speed, Amount of Gesticulation
Huang and Mutlu (2013)	Between-subject	Number of gestures, randomly selected	Iconic, Beat, Deictic, Metaphoric	Questionnaire + Retelling Performance	Naturalness, Competence, Effective use of Gestures

Salem, Eyssel, et al. (2013)	Between-subject	Unimodal (speech only), congruent multimodal, incongruent multimodal	Iconic, Deictic, Metaphoric	Questionnaire	Human likeness, Likability, Shared Reality, Future Contact Intentions
Salem et al. (2012)	Between-subject	Unimodal versus multimodal (speech + gestures) in a kitchen task	Iconic, Deictic, Metaphoric	Questionnaire	Gesture Quantity, Gesture Speed, Gesture Fluidity, Speech-Gesture Content, Speech-Gesture Timing, Naturalness
H. Kim et al. (2012)	Within-subject	-	Deictic, Beat	Questionnaire	Suitability of Gestures, Synchronization, Scheduling
Q. A. Le and Pelachaud (2012)	Within-subject	Synchronized Gestures, not Synchronized Gestures, Gestures with Expressivity, Gestures without Expressivity	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Synchronization, Naturalness, Expressiveness, Contradictiveness, Gestures are complementary, Gesture-speech Redundancy
Q. Le et al. (2012)	Within-subject	One Condition	Iconic, Beat, Deictic, Metaphoric	Questionnaire	Speech-Gesture Synchronization, Expressiveness, Naturalness
Salem et al. (2011)	Between-subject	Study 1: Unimodal versus Multimodal; Study 2: Same	Iconic, Deictic, Metaphoric	Questionnaire	Appearance, Naturalness, Liveliness, Friendliness
Levine et al. (2010)	Within-subject	Generated versus Ground Truth	Iconic, Beat	Pairwise	-
Ng-Thow-Hing et al. (2010)	Within-subject	4 studies: Audio vs Wrong Audio; Excited vs Calm Gestures; Low Expressivity, Medium Expressivity, High Expressivity; Slow Gesticulation, Medium Gesticulation, Fast Gesticulation	Iconic, Beat, Deictic, Metaphoric	Pairwise	-
Neff et al. (2008)	Within-subject	Speaker 1, speaker 2	Beat, Metaphoric	Match style to speaker	-
Bennewitz et al. (2007)	-	-	Iconic, Beat, Deictic	Public Exhibition	-

Table 2.3 Subjective Evaluations

Additional Results – Beat Gestures

Research work that focused on *only* beat gesture generation was excluded from the main analysis. Methods used to evaluate the performance of beat gesture generation systems in ECAs were similar to those used in work on semantic gesture generation. 10 papers were selected that met the criteria (Bremner, Pipe, et al., 2009; Chiu & Marsella, 2014; Fernández-Baena et al.,

2014; J. Kim et al., 2012; Kipp et al., 2007; Kucherenko et al., 2019; Levine et al., 2009; Ondras et al., 2021; Takeuchi et al., 2017; Wolfert et al., 2019). A total of 7 (70%) studies mentioned the number of participants, with a total of 236 participants. Only 4 (40%) mentioned statistics on age and gender. Of the 10 studies, 4 (40%) were performed in a lab, and 5 (50%) online or via AMT. 1 study was evaluated in an exhibition. As beat gesture generation mostly relied on prosody information, 8 (80%) studies used a data-driven approach. Only 4 of the 8 studies that relied on data-driven methods reported their metrics used for an objective evaluation, with either the Average position error (APE) or the MSE. 7 (70%) of papers ran their evaluation on a virtual avatar or stick figure with no discernible face. The subjective evaluations performed in these studies were similar to studies that included more gesture categories. 6 (60%) used a post-experiment questionnaire to assess the quality of the generated gestures by the ECA. 30% relied on pairwise comparisons and 1 (10%) relied on the time spent with focused attention on an ECA (Bremner, Pipe, et al., 2009). All studies (n = 10) relied on a within-subject evaluation. The questionnaire items that were used the most: 'naturalness' (n = 4, 40%) and 'time consistency' (n = 4, 40%).

2.8 Recommendations for Gesture Evaluation

As our thesis is concerned with mostly subjective evaluations, we looked at the usage of subjective evaluations by other researchers over the past years. Here, we provide recommendations that follow from reviewing the existing body of work, and we also use these recommendations in our next chapters.

2.8.1 Participant Sample

Many studies failed to report on the details of the participant samples. Additionally, not all participant samples reflect the data on which models or systems are trained. We recommend subjective evaluations with participants from diverse populations and backgrounds, reflecting the data on which models or systems are trained.

Some work is more focussed on equipping virtual agents with gesticulation, whereas others take it a step further and use their methodology to drive non-verbal behavior in social robots. Often, intermediate evaluation is overlooked, which can potentially lead to unwanted results when these engines are used in an interactive scenario. We recommend that participant evaluation is conducted -when feasible- before putting the model in production or when using the model on a new data-set, ensuring better validity and relevance when deployed for human social interaction.

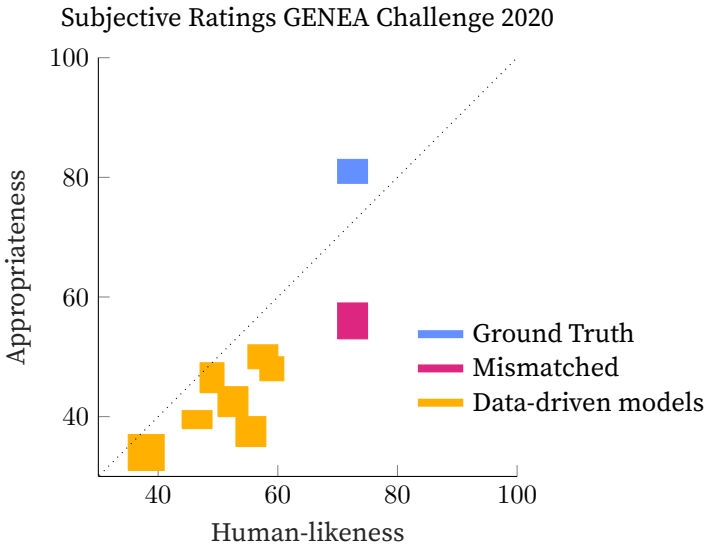


Figure 2.2 Human-likeness and appropriateness subjective measurements comparisons between data-driven models and the ground truth from the GENEA 2020 Challenge. Adapted from (Kucherenko, Jonell, Yoon, Wolfert, & Henter, 2021).

2.8.2 Experimental setup

The cornerstone of each subjective evaluation is to compare the output of a system to the ground truth. This ground truth condition must contain both motion and audio. Another condition that can shed light on a system's performance, is a random or mismatched condition, in which real motion is put on top of a different audio track. An interesting example of this is the subjective evaluation that was part of the Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA) 2020 Challenge, part of the International conference on intelligent virtual agents (IVA), and to our knowledge, the first of its kind in this field (Kucherenko, Jonell, Yoon, Wolfert, & Henter, 2021). In this challenge, multiple data-driven co-speech generators were compared to two baseline systems. A crowd-sourced subjective evaluation was part of this challenge, for which the results on 'appropriateness' and 'human-likeness' are displayed in Figure 3. Here, we see that ground truth is scored higher than the submitted systems on both dimensions and can function as a proper baseline. As for human-likeness, the mismatched condition offers an intriguing result: it does still look as human-like as the ground truth, yet it is scored much lower on appropriateness. Both a ground truth condition and a mismatched condition (i.e. where the visible motion does not match the underlying audio track) can function as a sanity check when being compared to the output of a system.

Most studies that we analysed asked participants to rate individual stimuli.

This can be substantiated with more rigor using the contrastive approach, also known as A/B testing or side-by-side testing (Kohavi & Longbotham, 2017). With such an approach, two or more stimuli are presented at the same moment, and a user is asked to either rate both stimuli or to select the preferred stimulus. In a recent study by the authors (which can be found in section 3.2), these two types of a contrastive approach were tested, as we wanted to find out whether one of the two contrastive approaches should be preferred (Wolfert et al., 2021). In one condition, participants were asked to make a choice between two videos (pairwise comparison) or to rate both videos. The authors found that when evaluating many conditions, an approach that makes use of rating scales is to be preferred over using pairwise comparisons. However, pairwise comparisons are a lot faster and less cognitively demanding on participants (Weijters et al., 2010).

Many studies evaluate the performance of their approach in a one-way fashion: videos are put online and participants are asked to evaluate individual videos. However, the need for proper gesticulation in ECAs is often tied to how humans communicate with each other. We recommend (when feasible) evaluating these systems in an interactive scenario, given that it is often the aim of researchers to eventually use ECAs in interactive scenarios. This might require additional engineering, such as creating systems that can also deal with synthetic speech (and thus with entirely new input), and creating dialogues to be used in an interactive scenario. However, by using an interactive scenario to evaluate an ECAs performance, it becomes possible to record and annotate interactions for indirect measurements, which we will discuss in the next paragraph.

A common way of evaluating stimuli is to ask for ratings on certain dimensions on a 5 or 7 point scale. Table 2.3 shows us the richness in terms of questionnaire items used for subjective evaluations. These items can also be seen as 'direct' items since they are used for direct measurement on a certain dimension, and cannot be captured by objective metrics (i.e. automated measures). Frequently used items for this are 'naturalness', 'human-likeness', 'appropriateness', or 'likability'. Our recommendation here, when one wants to rely on direct measurements only, is that subjective evaluations cover specific dimensions: naturalness, human-likeness, fluency, appropriateness, or intelligibility. Ideally, these dimensions are scored on a 5 or 7 point scale (as these tend to provide more reliable results than larger scales (Dawes, 2008)). In addition to direct measurements, we would like to make the case for using a more indirect way of measuring. Examples of indirect measurements are the time it takes to complete a task (task completion), recall rate (recall of facts when letting an ECA tell a story), eye contact and gaze, or response duration (in a question-answering session). For example, task completion is an often-used proxy to estimate effectiveness in human-computer interaction (Jordan, 2020), and might serve a similar role in our domain. The recall rate has already been used to evaluate gestures (Huang & Mutlu, 2013, 2014), but could play a more important role in future interactive evaluations. Eye contact, gaze, or response duration are good proxies to estimate a user's engagement, and tak-

ing engagement into account has worked well for other domains (Lemaignan et al., 2016; Nakano & Ishii, 2010). The level of engagement could in turn be a good predictor of how effective an ECA's gesticulation is. However, the drawback of using indirect ways of measuring, is that some of these approaches require annotating video recordings of experimental sessions with multiple raters.

2.8.3 Qualitative Analysis of Model Output

Data-driven models are often trained on a combination of speech-audio and text. Whereas some systems rely on one speaker (as is the case with systems submitted for the GENEA 2020 Challenge), others rely on multiple speakers. When data-driven systems are capable of generating gestures independent of a specific input voice, it becomes possible to use synthetic text-to-speech as input. This in turn makes it possible to present new data and to qualitatively analyze the performance of models on this new data. We propose a new task that takes entirely new sentences (and text-to-speech output when necessary) as input for gesture generation models. The output then needs to be analyzed for the occurrence of gesture categories. For example, for the sentence "I was throwing a ball", a model might generate an iconic gesture for the word 'ball'. We have crowdsourced a set of sentences and scenarios that can be used for this task [1]. We propose that researchers take a subset of these as input and that they annotate the model's output for the occurrence of gesture categories. This approach can provide an insight into the richness and diversity of the output of these models. However, this task only works for systems that can work with either only input text or a combination of input text and synthetic speech audio.

2.8.4 Preferred reporting items for Gesture Generation Researchers

To supplement the recommendations made in the previous sections, we offer a non-exhaustive list with preferred reporting items. These draw upon our observations of reporting and our research experiences (Kucherenko, Jonell, Yoon, Wolfert, & Henter, 2021; Wolfert et al., 2019, 2021). Considering the items in the proposed list, researchers could further enhance the quality of their reporting. Our proposed list with items that would be worth including in future work is summarized in Table 2.4. It contains items we deem important to report in a scientific publication when working on gesture generation for both physical and non-physical agents. We hope that the use of this list will make it easier in the future to allow for more systematic evaluation and benchmarking.

[1] https://github.com/pieterwolfert/gesturegeneration-checklist

Table 2.4 Preferred reporting items for co-speech gesture evaluation

Embodied Conversational Agent:
☐ ECA: Avatar/robot
☐ DOF (shoulder, elbow, wrist, hand, neck)
☐ Level of articulation of hands
Demographics:
☐ Recruitment method
☐ Sample size
☐ Age
☐ Gender distribution
☐ Geographical distribution
☐ Prior exposure with ECAs
☐ Language(s) spoken
Gesture Generation Model:
☐ Included generated gestures: [iconic, metaphorical, beat, deictic]
☐ Gesture generation model: [rule based, data driven, both, other]
☐ Gesture generation model link/repository
☐ (If not included – why not?)
Gesture Generation Evaluation:
☐ Context / application
☐ Evaluation method/questionnaire set
☐ Gestures annotated by human raters? [Yes/No]
☐ How many human raters were used?
☐ Inter-rater agreement
Metrics:
☐ Objective metrics [average jerk, distance between velocity histograms]
☐ Subjective metrics [human likeness, gesture appropriateness, quality, other]
Training dataset:
☐ Domain of dataset
☐ Length/size of dataset
☐ Gesture types annotated in the dataset
☐ Details on the actors in the dataset (N, language, conversation topic)
Statistical analysis scripts:
☐ Link to scripts

2.9 Conclusion

In this chapter we covered the relevant background for further reading of the thesis. We looked at nonverbal behaviour, and why generating it is important. We covered the body of work that has been done in the field of nonverbal behaviour generation, and more specifically, gesture generation for ECAs.

We discussed objective metrics and how they are calculated, which is necessary for understanding the objective evaluations discussed in later chapters.

We also reviewed the existing field specifically on subjective evaluations, as subjective evaluations are a key part of this thesis. In addition to our review, we provide recommendations for the field, on how to report and improve on methods and metrics.

Comparing and Evaluating Gestures

3

Comparing and Evaluating Gestures

In the previous chapter, the background for this thesis and related work was introduced. We found that in many cases, previous work does not include comparisons against other baselines, or uses standardised evaluation methods for evaluating computer generated nonverbal behaviour. In this chapter we look at several ways of evaluating computer generated speech behaviour for ECAs. In section 3.1 we look at the comparison of computer generated beat gestures with hand crafted beat gestures, with three different conditions for hand crafted gestures. We take things a step further in section 3.2, where we compare rating stimuli on a scale from 0 to 100 with the method of pairwise comparisons.

Since there are circumstances under which these direct measurements provide too little information, section 3.3 looks at the evaluation of a questionnaire for generated nonverbal behaviour.

3.1 Beat Gestures and Ranking

We wanted to know whether participants would be able to distinguish between hand crafted beat gestures and computer generated beat gestures. Because, if participants would not see strong differences, we would not have the need for generating new beat gestures, or we could use simpler, more computational efficient approaches, to drive ECAs. To explore this, we compared the output of a machine learning model with hand crafted beat gestures. For the machine learning part, we used the model by Kucherenko et al. (2019), which was trained on the Trinity dataset, featuring an English-speaking Irish actor (Ferstl & McDonnell, 2018). For the hand crafted beat gestures, we designed three baseline conditions: designed beat gestures, timed beat gestures and noisy gestures (not specifically beat gestures).

3.1.1 Problem Formulation

For the machine learning model, we considered the task of learning a mapping from a human speech signal to the corresponding upper body motion sequence: $\mathbf{m} = F(\mathbf{s})$, where $\mathbf{s} = (s_1, s_2, ...s_t)$ is a sequence of the prosodic

features from the speech signal and $\mathbf{m} = (m_1, m_2, ...m_t)$ is a sequence of 3D positions of the joints of a human skeleton. We describe speech features and motion joints below.

Speech Features

As speech features, we used four prosodic features, extracted with a window length of 5.55 Milliseconds (ms), resulting in 180 Fps, which were subsequently sub-sampled by averaging to 60 Fps. Those four features were: the energy of the speech signal, the logarithm of the F0 (pitch) contour and their numerical derivatives. The pitch and intensity value were extracted from audio using Praat (Boersma & Weenink, 2002) and normalized as in Chiu and Marsella (2011).

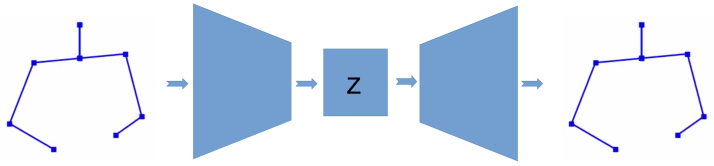
Human Skeleton

Since our focus was on beat gestures, we only took the upper body (excluding the fingers) into account, when generating gestures. This resulted in a skeleton containing 8 joints: head, neck, left shoulder, right shoulder, left elbow, right elbow, left hand, and right hand.

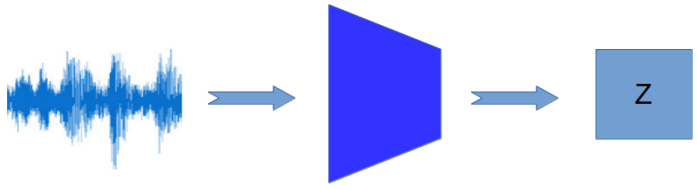
3.1.2 Deep-Learning Based Solution

The machine learning model for speech-driven gesture generation we used for this, is depicted in Figure 3.1. It is composed of two parts (Figures 3.1a,b), which are then combined together (Figure 3.1c). First, a lower-dimensional representation of human motion is learned using a Denoising Autoencoder neural network. This neural network consists of a motion encoder *MotionE* and a motion decoder *MotionD*. Second, another neural network *SpeechE* is trained to map from speech to a corresponding motion representation. At test time, the speech encoder and the motion decoder networks are combined: *SpeechE* predicts motion representations based on a given speech signal and *MotionD* then decodes these representations to produce motion sequences.

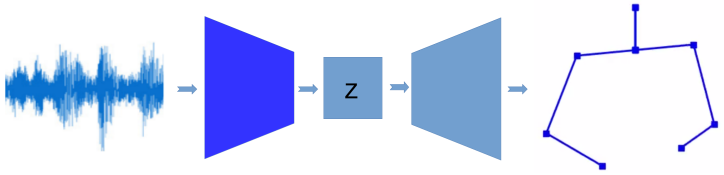
Since the human skeleton considered in our experiments is much simpler than the one in the original paper and the dataset is significantly smaller, the network was significantly reduced. For the Denoising Autoencoder (Figure 3.1a) the representation dimensionality was 20 instead of 325. The speech-to-representation neural network (Figure 3.1b) was also significantly simplified: the hidden layer size was reduced to 36; the amount of layers to 3: Fully Connected layer, Gated Recurrent Unit, and another Fully Connected layer; the batch size was reduced to 128; and the initial learning rate to 0.0005.



(a) MotionED: Representation learning for the motion



(b) SpeechE: Mapping speech to motion representations



(c) Combining the learned components: SpeechE and MotionD

Figure 3.1 How the encoder-decoder Deep neural network (DNN) for speech-to-motion mapping (Kucherenko et al., 2019) is constructed. Every trapezium denotes a neural network, z denotes encoded representation of motion.

3.1.3 3D Upper Body Modelling

For manual beat gesture generation, a 3D model of a human's upper body was modelled (with the joints as specified in section 3.1.1) using the Unified robot description format (URDF). URDF is a XML file format in which joints, dimensions and links are specified, and therefore this file can describe the kinematic information of the described agent. URDF can be used for running simulations with Robot operating system (ROS) (Quigley et al., 2009). Having this URDF file that described our 3D model of the human upper body, we were able to calculate inverse and forward kinematics with the Python Module IKPY [1]. This served as the basis for our modelled beat gestures.

3.1.4 Experimental Design and Conditions

Dataset

We trained the model on the Trinity College Conversational Dataset (Ferstl & McDonnell, 2018). This dataset features one actor speaking freely about a variety of topics. Together with the video and audio, the motion of the actor was captured using Motion capture (MoCap) system. In total there are 23 takes of roughly 10 minutes.

Generated Beat Gestures

Machine Learning Generated (Condition 1) For the first condition we fed the trained model 10 seconds of audio, and concatenated and smoothed the resulting pose positions. To fit the generated poses in the same frame as the other pose positions in other conditions, we normalized and post-processed the resulting skeletons such that the location of the neck was at (0, 0, 0). The poses were also rotated to make the resulting skeleton facing front. [2]

Designed Beat Gestures (Condition 2) For our manual gesture conditions, we used a 3D model of the human upper body, as described in section 3.1.3. The start position was with the hands in a resting position, where the hands are close to the hips. To generate beat gestures, we applied a vertical translation from the average resting position (McNeill, 1992). The trajectory is defined with a sine function on the y-axis. The amplitude of the sine function was alternated to generate different types of beat gestures. To make sure that the trajectories of the hands appeared natural, for every position in Cartesian space new joint positions were calculated through inverse kinematics, hence the need for a 3D model. To arrive at natural looking gestures, the x-values for the sine function were drawn from a logarithmic scale, from zero to π.

[1] github.com/Phylliade/ikpy

[2] C1: https://youtu.be/AJlc54yODPw, C2: https://youtu.be/I5c3FgWgdjY, C3: https://youtu.be/ONehBn8N9a8, C4: https://youtu.be/bXUS3SQBg9w

Beat gestures with different amplitudes were concatenated at random, and combined with audio. To smooth the concatenation of gestures, the input joint positions of a new gesture were the last known joint positions of the previous gesture.

Timed Beat Gestures (Condition 3) For the third condition, noisy gestures were sampled (generation of these gestures was similar to that of condition two, but with a very small amplitude). On top of these noisy gestures, a beat gesture was added roughly 400 milliseconds before a pitch in the audio was detected Kendon, 1980. The onset of several pitches were taken, and the loudest pitch was taken as the pitch to input a beat gesture. Pitch detection and other on-the-fly audio processing was done using Librosa McFee et al., 2015.

Noisy Gestures (Condition 4) Noisy gestures were generated like our designed beat gesture generation, but with a very small amplitude (to resemble noise on the endpoints). As these are context free, i.e. no speech input is used for the timing and they were not designed to resemble human-like beat gestures, the prediction is that this condition will be ranked lowest.

3.1.5 User Study

We used 10 audio samples of 10 seconds, from which we generated 40 videos of 10 seconds, which in turn translates to a video per condition, per audio file. To run pairwise comparisons, we needed in total six pairs per sample, which brought the total number of comparisons to sixty. For every pair, the user was asked to select the video which had the users' preference. A survey was set up using Google Forms, and the order of conditions was counterbalanced, to minimize the chance that two of the same conditions would succeed each other in the survey. The survey was promoted through Amazon Mechanical Turk. To control for the worker's focus, we added a control question, and logged the amount of time it took to complete the survey. Surveys completed in less than 10 minutes were not seen as serious submissions, and not taken into account for our analyses.

3.1.6 Results

We evaluated and compared the different conditions with each other through a user study. 41 participants were presented 59 video pairs[3]. Of these 41 participants, the average age was 33 years (SD=9.5 years). Nineteen of them were male, twenty-two of them were female. 40 participants were native English speakers, 1 was not.

Since we used pairwise comparisons, we ran a Chi-Square Goodness of Fit test per pair (six in total). For this test, we assumed that if the conditions would be

[3]Due to an error on our side one pair was left out the survey.

ranked equally, the distribution would be 50/50 per pair. For all six possible pairs, $p < 0.05$, and this assumption was therefore rejected.

A ranking was deduced using the Bradley terry luce (BLT) model (Bradley & Terry, 1952). This method is used to calculate a ranking based on pairwise comparisons. In this situation we decided not to let participants rank all conditions at the same time, but compare them against one and other. The BLT model provides a prediction p for the outcome of a paired comparison, where this prediction is in the form of the logarithm of the odds, $log(\frac{1}{1-p})$. Logarithm of the odds is a method to map $p[0, 1]$ to $[-\infty, +\infty]$, where a logit less than 0 equals $p < 0.5$. The results of applying this model to our data can be found in table 3.1. Given the results a global ranking of our conditions from the preferences of the users in our user study can be derived, as visible in Figure 3.2, where we can see the number of 'votes' per condition (following the results from the ranking model).

Table 3.1 Logit of Winning

	Wins			
Losses	Condition 1	Condition 2	Condition 3	Condition 4
Condition 1	-	0.94	1.81	2.23
Condition 2	-0.94	-	0.88	1.30
Condition 3	-1.81	-0.88	-	0.42
Condition 4	-2.23	-1.30	-0.42	-

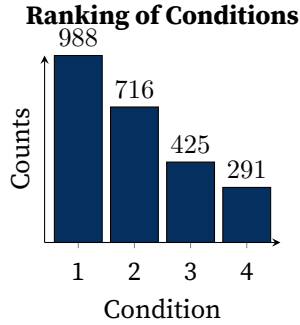


Figure 3.2 Ranking based on number of wins (max is 1230).

3.2 Ratings versus Comparisons

For this section, we compared two different evaluation approaches for computer generated nonverbal behaviour. Although similar comparisons have been done in other fields, it had not been done in relation to the evaluation

of gesture generation for ECAs. The first evaluation approach we look at is the use of rating scales for human-likeness. This approach for measuring human-likeness was introduced with the GENEA 2020 challenge, and also evaluated before the challenge by Jonell et al. (2021). The second metric, that has been used before as well (see chapter 2), is the method of pairwise comparisons. With pairwise comparisons, two stimuli are presented, and the participant is asked to choose one of the two comparisons (or say that both are equal). Relatively little empirical attention has been devoted to this methodological topic in regard to the evaluation of data-driven generated stimuli, however, it is still unknown how much the methods actually differ in terms of usability and informativeness.

3.2.1 How to measure?

There is a rich body of work in psychology on how to measure that what you want to measure. DeCoster et al. (2009) compared analysing continuous variables directly with analysing them after dichotomisation (e.g., re-coding them as two-class variables such as high-or-low). Although there were a few edge cases where dichotomisation was similar to direct analysis, they demonstrated that dichotomisation throws away important information and concluded that the use of the original continuous variables is to be preferred in most circumstances. Simms et al. (2019) randomly assigned participants to complete the same personality rating scales with different numbers of response options ranging from two to eleven. They found that including four or fewer response options often attenuates psychometric precision, and including more than six response options generally provides no improvements in precision. Finally, Rhemtulla et al. (2012) demonstrated that treating rating scale data as continuous can be problematic (i.e., can result in biased estimates) for scales with fewer than five response options, which tend to be quite non-normally distributed. Such data thus requires specialised ordinal methods to analyse properly. Overall, the psychological literature thus suggests that rating scales with between five and seven response options would be preferable to rating scales with fewer response options. If we consider the pairwise comparison approach to be similar to a rating scale with two response options (e.g., better or worse), this would raise concerns about the approach's psychometric precision and normality.

Another way of evaluating stimuli on a continuous scale is by using visually-aided rating (VAR) (Janhunen, 2012). Visually, categories are still used as anchors in VAR, but specific scores are not visualised in comparison to Likert scales. This enables participants to quantify an ordering, from which it is still possible to derive a quantifiable rating. VAS-RRP is congruent to VAR, except that in VAR the rating scale is placed vertically, and in VAS-RRP horizontally (Sung & Wu, 2018).

However, there have also been impassioned arguments in favour of ordinal and rank-based approaches (of which the pairwise comparison approach can

be considered a simple variant) within the affective computing community in recent years (Martinez et al., 2014; Yannakakis & Martínez, 2015; Yannakakis et al., 2021). Ranking is something we explored in section 3.1. The argument is that many subjective evaluations are inherently ordinal and cannot be adequately treated as continuous numbers or nominal categories and should instead be handled using rankings. If this argument is accurate, then the pairwise comparison approach would be preferable to the rating scale approach on theoretical grounds. There is also evidence that rank-based approaches might have some practical benefits over rating scale approaches, such as being faster to administer and more reliable over time. For example, Clark et al. (2018) evaluated the perception of physical strength from images of male bodies using both pairwise comparisons and rating scales and found that the scores were closely correlated but that the pairwise comparisons were completed 67% faster. Other examples, like Elliott (1958) and Mueser et al. (1984) found high correlations between rankings resulting from the evaluation of physical features in humans. Liang et al. (2020) proposes a model to 'calibrate' self-reported user ratings for dialogue systems due to issues with validity and bias. In relation to biomedical image assessments, where evaluation considers the visual quality of the stimuli, Phelps et al. (2015) found that pairwise comparisons and ranked Likert scores made for more accurate assessments in comparison to the use of non-ranked Likert scores. Burton et al. (2019) compared rating scales with best-worst scaling, another variant of the rank-based approach. In this study, participants were asked to select the most attractive and least attractive faces in a series of images. The best-worst scaling approach showed better test-retest reliability than the rating scale approach.

3.2.2 Hypotheses

To study the differences and comparisons between using a rating or pairwise comparison approach for the evaluation of generated co-speech gestures, we defined five hypotheses. Our hypotheses, design, and methodology were pre-registered before data collection[4]. We present short video clips to human participants, with each video clip showing an avatar displaying combined verbal and nonverbal behaviour. The movements are generated using three data-driven methods of varying quality and we expect the subjective evaluations to clearly reflect this difference. In order to gain more insight into the effectiveness of the two subjective evaluation methods, we formulated the following five hypotheses.

H1. The rank-order of stimuli implied by the pairwise comparisons and rating scales will be different.

H2. Pairwise comparisons will have higher inter-rater agreement than rating scales.

[4]https://osf.io/7d9fs

H3. Pairwise comparisons and rating scales will differ in terms of time-efficiency (e.g., the time it takes for a single participant to finish a single evaluation).

H4. Pairwise comparisons and rating scales will differ in terms of participant usage preference and usability (both qualitative and quantitative).

H5. Pairwise comparisons and rating scales will both find a difference between stimuli that have a pronounced quality difference, but will not have enough resolution to find a difference between stimuli that differ slightly in quality.

3.2.3 Experimental Design and Conditions

We used 30 video stimuli[5] showing a gesticulating avatar provided by Kucherenko et al. (2020), the stimuli are already available and have been used by other researchers (Jonell, Kucherenko, Torre, & Beskow, 2020). The videos had a resolution of 640×480 pixels and a frame rate of 30 frames per second. Three types of videos were used: Full, NoSpeech and NoText. The *Full* videos were generated by a model trained on motion of a human actor with the model having access to both the audio speech and transcribed text; the *NoSpeech* videos were generated from a model only trained on motion and transcribed text; and the *NoText* videos were generated by a model trained on motion and speech audio only. Thirty videos were created per type and, in each triplet of videos (across type), the avatar spoke the same sentence to facilitate comparison. We have two study conditions: Full versus NoSpeech (which we denote Low Difference) and Full versus NoText (High Difference). We denote them this way because the former showed a small difference in the original study (Kucherenko et al., 2020), while the latter showed a large difference. These conditions (Full. vs NoSpeech and Full. vs NoText) turned out to show significant differences in quality, and we assume that our subjective evaluations will reflect this.

Each participant was first assigned to either the LowDiff or HighDiff condition. Following that, the participant was assigned to one of two ordering conditions:

PR: Pairwise Comparison approach for 10 videos drawn from a set of 30 videos, followed by the Rating Scale approach for the same 10 videos.

RP: Rating Scale approach for 10 videos drawn from a set of 30, and then Pairwise Comparison approach for the same 10 videos.

[5] http://svito-zar.github.io/gesticulator/

Participants

We recruited 130 participants on Prolific[6]. To ensure data to have quality, participants had to be a native speaker of English, had at least a 90% approval rating on the platform, and had to have participated in at least 100 other studies on the platform. Participants were assigned to conditions using block randomisation in order to maintain balanced conditions.

Technical Setup

From Prolific, participants were forwarded to a web application to evaluate the stimuli. This application was based on HEMVIP (Jonell et al., 2021), which in turn was based on WebMushra (Schoeffler et al., 2018) but adapted to work with video files. Since two evaluation strategies were evaluated, there were two interface versions.

The pairwise comparison interface (Figure 3.4) displays two videos side by side, with three options for evaluation displayed below the videos. For all conditions, the question was: 'In which video are the character's movements most human-like?' The three response options were: *left*, *right*, and *equal*. Participants were able to play both videos at the same time, but it is not explicitly mentioned in the instructions. After the participants watched both videos and selected a response option, they could continue to the next page.

The rating scale interface (Figure 3.3) displays a single video at a time, with a rating scale displayed below. For all conditions, the question was: 'How human-like was the agent in this video?' Response options ranged from 1 to 5 and were labelled *not at all*, *slightly*, *somewhat*, *moderately*, and *extremely*. Videos could only be watched one-at-a-time, and participants were only able to advance to the next page when both videos had been played and rated.

Experimental Procedure

After participants were assigned to the task on Prolific, they were forwarded to the online evaluation system. Here, they were assigned an internal participant ID that corresponds to a configuration file containing the stimuli and order of stimuli to show to the participant, and when to run attention checks. Each participant evaluated a total of 22 video pairs. These 22 video pairs correspond to 10 videos evaluated in a pairwise comparison approach, and 10 in a rating style approach. Two of the 22 video pairs contained an attention check. The order of evaluation (pairwise comparison vs. rating approach) was based on the assigned ordering condition. The position of the attention checks in the series of evaluation pairs was randomised, and there were two types of attention checks: one in which the response option to select was provided visually and one in which it was provided acoustically.

[6]https://www.prolific.co/

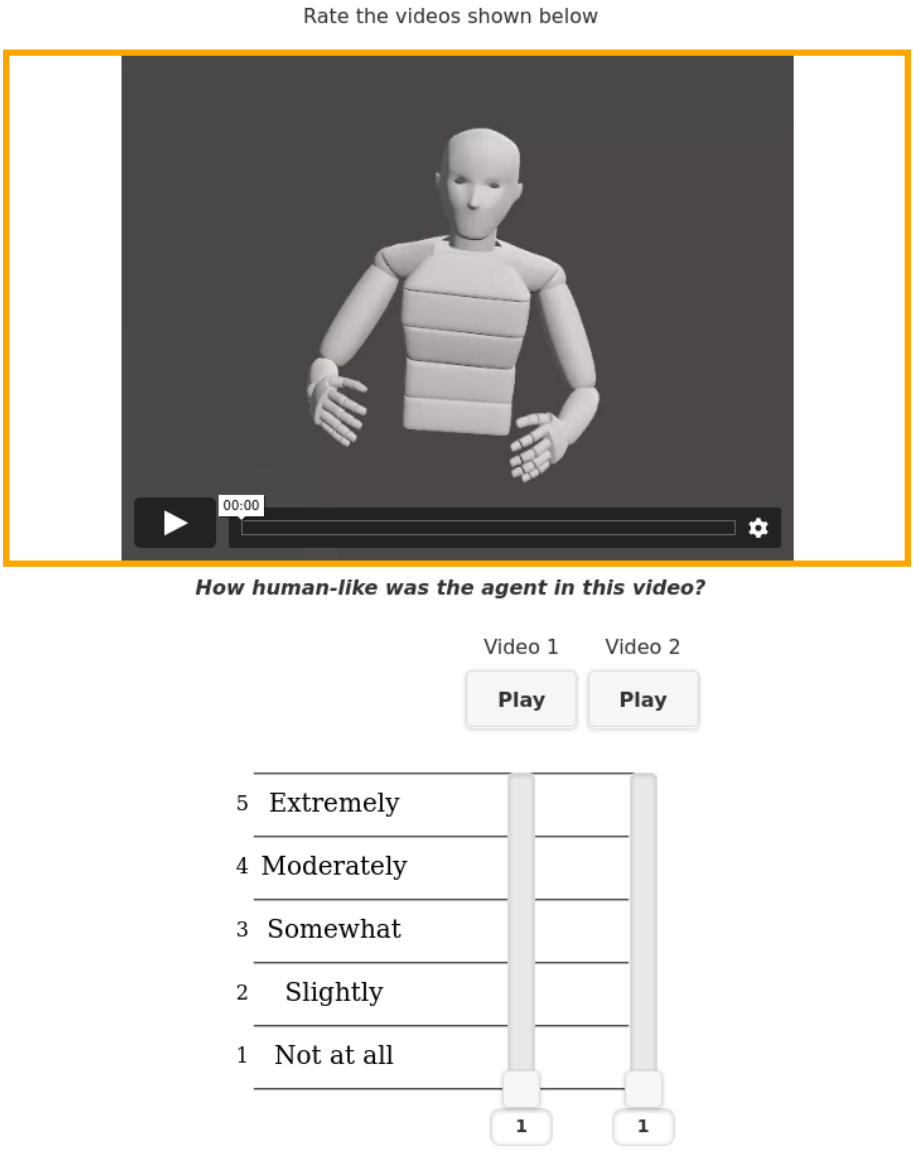


Figure 3.3 Interface for rating scale evaluation, note that the response options were 1 to 5, with an anchor for each option. This is different from the interface used in chapter 4 and 5 for the evaluation on human-likeness.

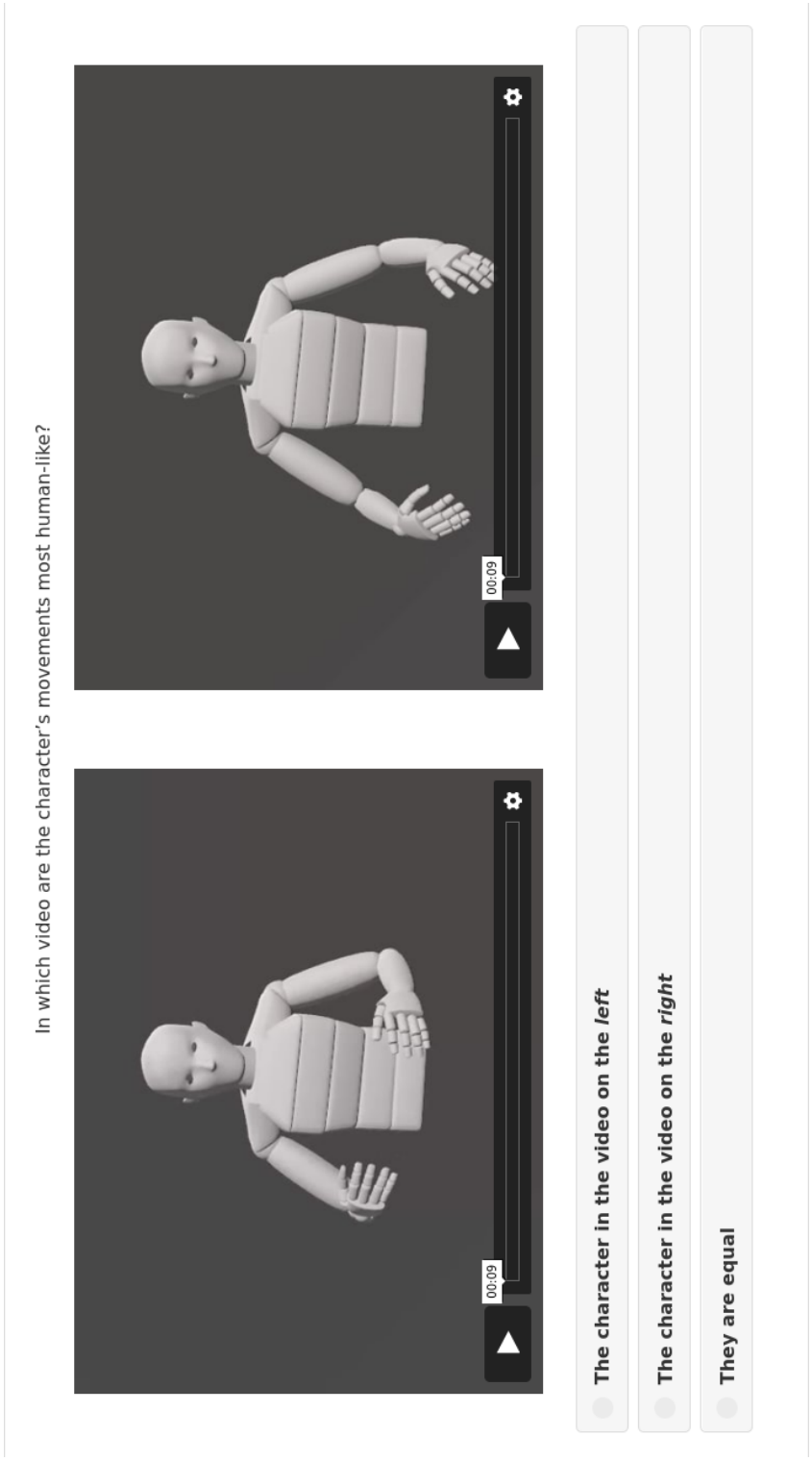


Figure 3.4 Interface for pairwise comparisons.

After evaluating the 22 video pairs, participants were presented with a questionnaire collecting their age, gender, nationality, level of education and experience with computers. This was followed with open questions related to the procedure they just completed, and whether they had a preference for pairwise comparison or rating scale evaluations. Once done with the study, successful participants were rewarded with 2.50 GBP (pay on average was 7.23 GBP per hour when taking into account the average duration of the task). The time each participant spent on each page of the experiment (and overall) was also recorded to allow us to evaluate efficiency.

3.2.4 Analyses

Hypothesis 1

To test the hypothesis that the two comparison methods would result in different rank-orderings of stimuli, we used a correlational approach. We first calculated each stimulus' average score across participants for each comparison method. Average scores using the rating scale method ranged from 1 to 5, and average scores for the pairwise approach ranged from –1 to 1 (on a scale where 1 = the stimulus was preferred over the alternative, 0 = the stimulus and alternative were equal, and -1 = the alternative was preferred over the stimulus). We then estimated the Kendall Rank-Order correlation (Kendall, 1938) between these two series.

Hypothesis 2

To test the hypothesis that the pairwise comparison method would have higher inter-rater agreement than the rating scale method, we used two statistical approaches. First, we estimated intraclass correlation coefficients (ICCs) using Model 2A (McGraw & Wong, 1996), for we had two random effects without interaction, and calculated the absolute agreement of the average of 12 participants (i.e., the minimum number of participants assigned to any comparison). We selected this specific model since not all raters are fixed, and thus a random selected sample of all available raters. This approach estimates the reliability of the average of multiple participants' responses (which is what is used to compare video-generating methods), but assumes that the data approximates a continuous distribution (which is not the case for the pairwise method). As such, we also estimated chance-adjusted categorical agreement using quadratic-weighted kappa coefficients (Gwet, 2014). This approach is overly pessimistic in this case because it estimates the reliability of a single randomly selected participant's response, but it has the benefit of not assuming continuous data. In both cases, 2000 iterations of non-parametric bootstrapping (Efron & Tibshirani, 1993) (with percentile-based confidence intervals and p-values) were used to compare the two approaches' inter-rater reliability.

Hypothesis 3

To test the hypothesis that the two comparison methods would differ in terms of time-efficiency (i.e., the time it takes a participant to complete a single comparison/page), we used a linear mixed effects modelling approach (Gałecki & Burzykowski, 2013). We estimated a model in which each page's completion time (in seconds) was regressed on a binary variable representing the comparison method. To control for practice and fatigue effects, we also regressed the completion time variable on a binary variable representing whether the comparison was during the first or second half of the experiment, and the method-by-half interaction effect to allow the difference between comparison methods to differ between the first and second half of the experiment. Finally, to account for the clustering/nesting of comparisons within participants and videos, we included random intercepts for these variables and used Satterthwaite's approximation (Kuznetsova et al., 2017) to correct model degrees of freedom for small clusters.

Hypothesis 4

To test the hypothesis that participants would be more likely to prefer the pairwise comparison approach than the rating approach, we estimated an intercept-only logistic regression model to predict a binary variable representing whether each participant preferred the pairwise comparison approach over the rating comparison approach. We then back-transformed the intercept to probability units and tested whether it was significantly different from an equal preference of 50%.

Hypothesis 5

To test the hypothesis that the two comparison methods (i.e. rating scale and pairwise) would both find a difference in the case of a large difference in the quality of generated behaviour (i.e., Full vs. NoText stimuli) but not in the case of a small difference in the quality of generated behaviour (i.e. Full vs. NoSpeech stimuli), we used a linear mixed effects modelling approach (Gałecki & Burzykowski, 2013). We estimated a model in which the choice for the Full stimuli was regressed on other (NoText or NoSpeech) and order.

3.2.5 Results

130 participants were recruited, of which 100 participants passed the attention checks. Of these, the mean age was 35.01 (SD=12.64), 55 identified as female, 45 as male. 68 of the participants were UK nationals, 22 were from the USA, 4 participants were Canadian, 2 Irish, 1 Australian, 1 Bulgarian, 1 Indian and 1 from New Zealand.

Hypothesis 1

In Figure 3.5, we can see the relationship between the average pairwise scores and the average rating scores. We quantified the magnitude of this relationship using Kendall's Rank-Order Correlation. When we excluded trials where the two stimuli being compared were rated as equally human-like, we found a rank correlation of 0.44, 95% CI: [0.32, 0.55], $p < .001$. When we included trials where the two stimuli being compared were rated as equally human-like and assigned a pairwise score of 0, this correlation became 0.46, 95% CI: [0.35, 0.57], $p < .001$. Thus, although the two methods did not have exactly the same rank-ordering of stimuli, their rank-orderings were positively correlated (i.e., similar) to a high degree.

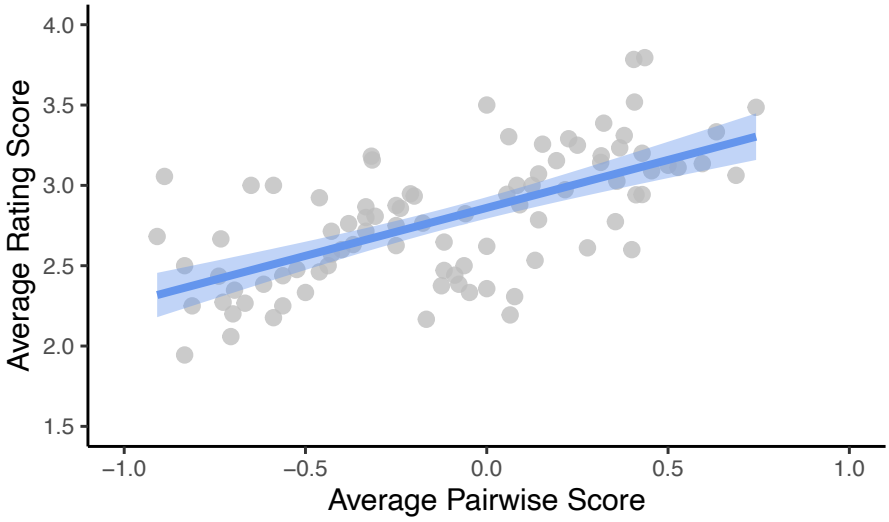


Figure 3.5 Relationship between average rating and pairwise scores. The two are positively correlated

Hypothesis 2

Using the intraclass correlation approach, the inter-rater reliability coefficient was 0.62, 95% CI: $[0.50, 0.69]$ for the rating scale method and 0.77, 95% CI: $[0.71, 0.82]$ for the pairwise method; this difference was statistically significant ($\Delta = 0.15$, 95% CI: $[0.06, 0.27]$, $p < .001$). Using the chance-adjusted categorical agreement approach, the quadratic-weighted kappa coefficient was 0.14, 95% CI: $[0.09, 0.18]$ for the rating scale method and 0.23, 95% CI: $[0.18, 0.28]$ for the pairwise method; this difference was statistically significant ($\Delta = 0.09$, 95% CI: $[0.03, 0.16]$, $p = .009$).

Hypothesis 3

The main effect of comparison method was significantly greater than zero, $B = 6.07$, 95% CI: $[2.36, 9.77]$, $p = .002$ (see Figure 3.6). The unstandardised slope estimate of 6.07 means that pages were completed an average of around 6 seconds faster for the pairwise approach than for the rating approach. The main effect of ordering was not significantly different from zero $(p = .491)$ and the type-by-ordering interaction effect was also not significantly different from zero $(p = .600)$, which means that completion time did not significantly differ between the first and second half of the experiment and that the difference between comparison methods did not depend on which came first or second in the experiment.

If we want to know what the time difference would be for an entire experiment, we can multiply this page-level effect by the number of pages shown to participants. For 10 pages, as we did in this study, the experiment-level difference would be around 60 seconds.

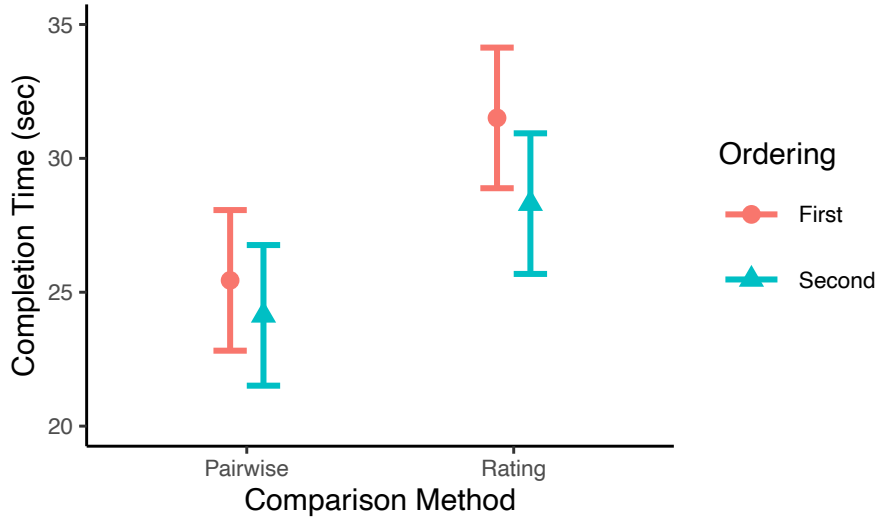


Figure 3.6 Completion time across conditions (error bars are 95% CIs), showing that the pairwise method is approximately 6 seconds faster per page than the rating method

Hypothesis 4

The intercept for preference for the pairwise method was estimated at 56.0%, 95% CI: $[46.2\%, 65.5\%]$ and was not significantly different from an equal preference of 50% $(p = .231)$. Thus, we cannot conclude that participants reliably preferred one method over the other.

Hypothesis 5

For the rating scale method, the main effect of other was significantly greater than zero, $B = 0.66$, 95% CI: $[0.40, 0.92]$, $p < .001$. This means that the extent to which the Full stimuli were rated higher than the other stimuli was greater for the HighDiff stimuli than for the LowDiff stimuli. In this model, neither the main effect of ordering ($p = .439$) nor the other-by-ordering interaction effect ($p = .860$) were significant. For the pairwise method, the main effect of other was significantly greater than zero, $B = 1.07$, 95% CI: $[0.45, 1.69]$, $p < .001$. This means that the probability of preferring the full stimulus over the other stimulus was greater for the HighDiff stimuli than for the LowDiff stimuli. In this model, neither the main effect of ordering ($p = .750$) nor the other-by-ordering interaction effect ($p = .094$) were significant. Despite different scaling, the two methods had very similar results that matched our hypotheses and also matched the results from the original study we were reproducing (Kucherenko et al., 2020) (see Figure 3.7).

3.3 Questionnaire Creation and Evaluation

In section 2.8.4 we provide pointers for a questionnaire and evaluation design, after we discussed the evaluation methodologies that have been used up till now. We identified that a majority of studies make use of questionnaires to assess the quality of the generated motion. Questionnaires, employing Likert scales, are a widely used tool to assess ones attitude towards a concept (Weiss & Bartneck, 2015). One such example is the Godspeed questionnaire (Bartneck et al., 2009), which originates from and is used in the field of human-robot interaction, and measures the concepts of anthropomorphism, animacy, likeability, perceived intelligence and perceived safety. Hence our call for standardisation given the large variety in reported constructs for the field of gesture generation (that consist of multiple items). In support of that, a 2019 review by Fitrianie et al. that looked into questionnaire usage at the intelligent virtual agent conferences (IVA), found that for 76% of the studies, questionnaires were unique for that study and were not reused in other studies (Fitrianie et al., 2019). Having matching constructs and questionnaire items would make it easier to compare between studies. Efforts for standardisation are underway for the evaluation of virtual agents, but these efforts do not specifically include the evaluation of nonverbal generated behaviour for ECAs (Fitrianie et al., 2020). Among the constructs that they report, are human-likeness, appropriateness, naturalness or effectiveness and understanding. Even though there is a degree of overlap of constructs between studies, this does not mean that they are measured in the same way and that they contain the same items (statements) and response scales.

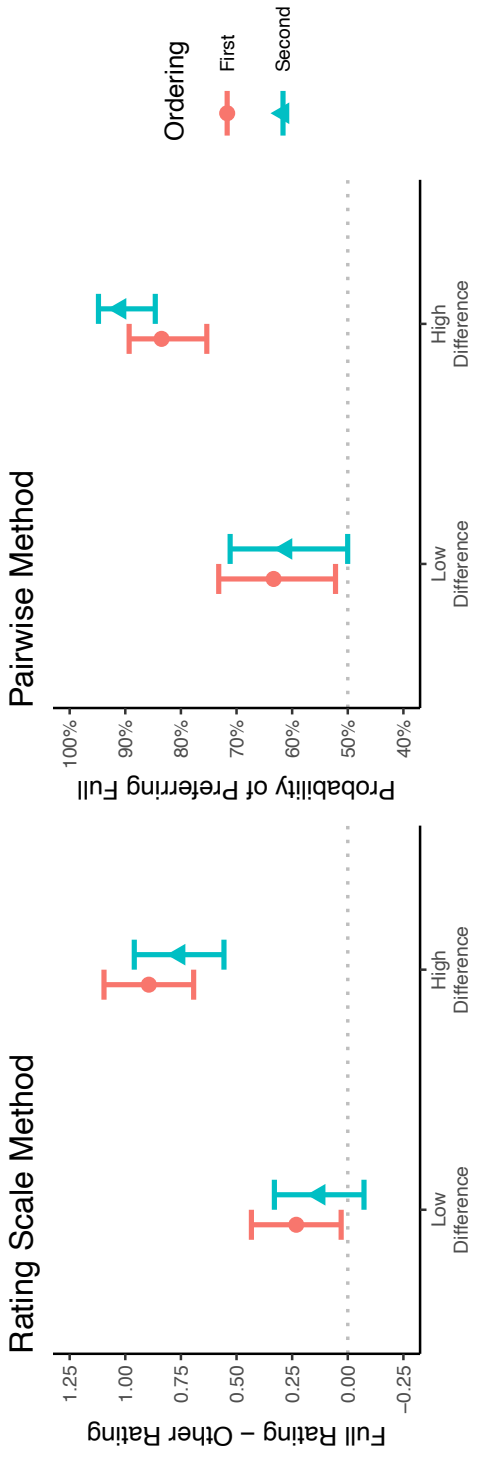


Figure 3.7 Comparison of generation methods by condition and evaluation method (error bars are 95% CIs)

3.3.1 Questionnaire Creation

We put forward a questionnaire that instead covers the concepts of appropriateness, human-likeness and intelligibility, which are key foci of nonverbal behaviour generation for the ECA community. The proposed questionnaire consists of three Likert scales with 5 Likert items each, and can be found in Figure 3.9. Wolfert, Robinson, and Belpaeme (2022) identified multiple studies that included questions on the appropriateness or speech-gesture correlation. The GENEA Challenge 2020 used a direct question related to appropriateness. We decided to include the construct of 'appropriateness', with 5 Likert items (statements) related to the appropriateness of the motion behaviour for the conversation. Since the concept of human-likeness often comes back in subjective evaluations and has been also used for direct questioning by the GENEA Challenge, we came up with 5 Likert items related to human-likeness of the gesture motion. Lastly, we want to evaluate the intelligibility of the agent/speaker motion, as this does also regularly appear in subjective evaluations of synthesised gesture motion (such as 'content' or 'utilisation of gesture' per Table 3 in the work by Wolfert, Robinson, and Belpaeme (2022)).

3.3.2 Questionnaire Evaluation

We make use of the questionnaire as described in the section on questionnaire design, for rating stimuli displaying both gesticulation (study 1) and listening behaviour (study 2). We recruited 46 participants for study 1 and 48 participants for study 2. Participants were asked to rate their level of agreement with each statement on a scale with 5 answer options: disagree, slightly disagree, neutral, slightly agree and agree. Each participant is presented with 8 videos, one per screen, and 1 attention check. The attention check consists of one statement of the 15 on a page that asks the participant to select one answer option. Each video is accompanied by 15 statements on the page for which the participant has to indicate their level of agreement. For this, we adapted the HEMVIP interface to display one video with 15 statements. Only upon answering all statements, the 'next' button would be activated. The interface is visible in 3.9.

3.3.3 Results

Gesturing

We examined the perceived quality of synthesised gestures across three dimensions: appropriateness, human-likeness, and intelligibility. To assess the internal consistency of the rating scales used for these dimensions, we calculated Cronbach's alpha coefficients. 46 participants (22 males, 24 females) participated and passed the attention checks. Of these participants, the mean

Appropriateness of the motion

- The motion seemed appropriate for the context of the conversation.
- The motion felt out of place or irrelevant to the interaction.
- The motion did not distract from the conversation.
- The motion was in sync with the pace of the interaction.
- The motion was in synchronization with the agentś tone of voice and emotion or their active listening.

Human-likeness of the motion

- The motion did not look like it was produced by a human.
- The motion appeared smooth and effortless.
- The motion had the same characteristics as human motion.
- The motion seemed forced or robotic.
- The speed of the motion looked human-like.

Intelligibility of the agent

- The motion enhanced the understanding of the interaction.
- The motion well captured what the character was trying to express.
- The meaning of the motion was easy to interpret.
- The motion helped me understand what the person was saying or showed that he was actively listening.
- The motion added to the perception of the agent's strong communication skills.

Figure 3.8 Participants were asked to rate each statement in the questionnaire on a scale from 1 to 5, using the following anchors: (1) Disagree, (2) Slightly Disagree, (3) Neither Agree nor Disagree, (4) Slightly Agree, (5) Agree.

age was 42 (SD=14.6). 40 participants resided in the UK, 1 in the USA and 4 in Canada.

For the appropriateness dimension, the Cronbach's alpha coefficient was 0.90 (95% CI [0.89, 0.92]), suggesting good internal consistency among the items assessing appropriateness. The human-likeness dimension had a Cronbach's alpha coefficient of 0.92 (95% CI [0.91, 0.94]), indicating high internal consistency among the items measuring human-likeness. Furthermore, the intelligibility dimension exhibited excellent internal consistency, as indicated by a Cronbach's alpha coefficient of 0.97 (95% CI [0.97, 0.98]). This suggests a high

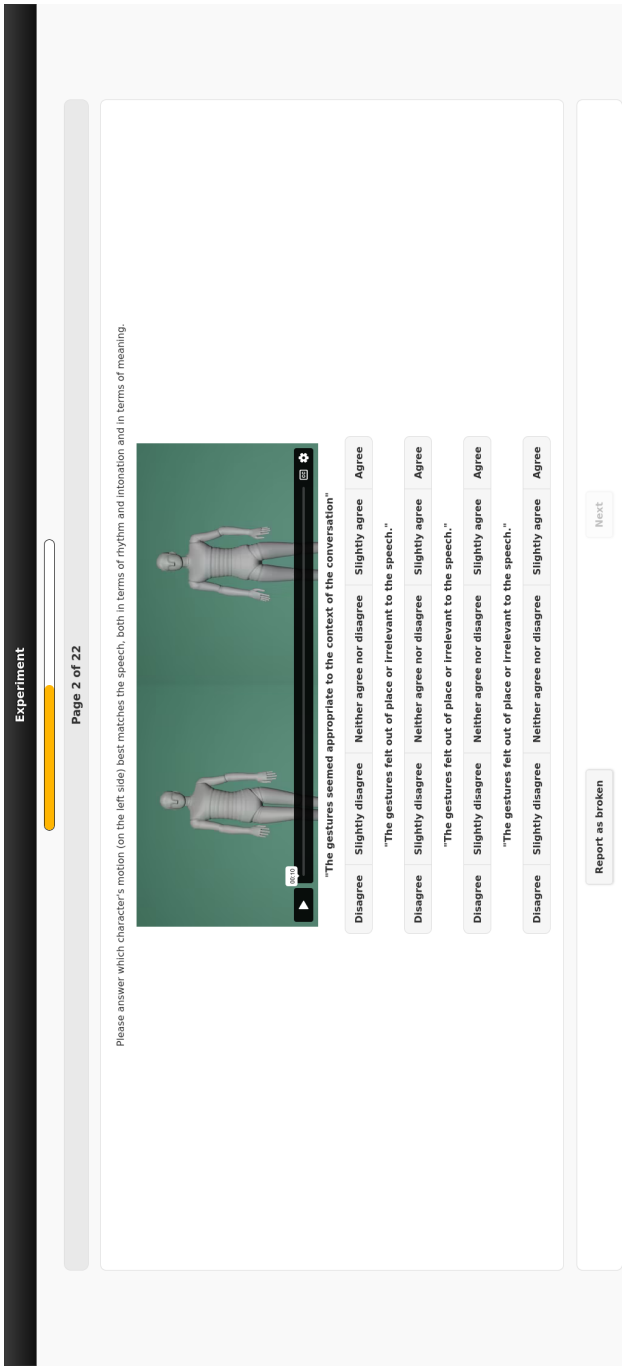


Figure 3.9 Interface used for questionnaire evaluation. Two avatars in conversation (with frontal view) are visible.

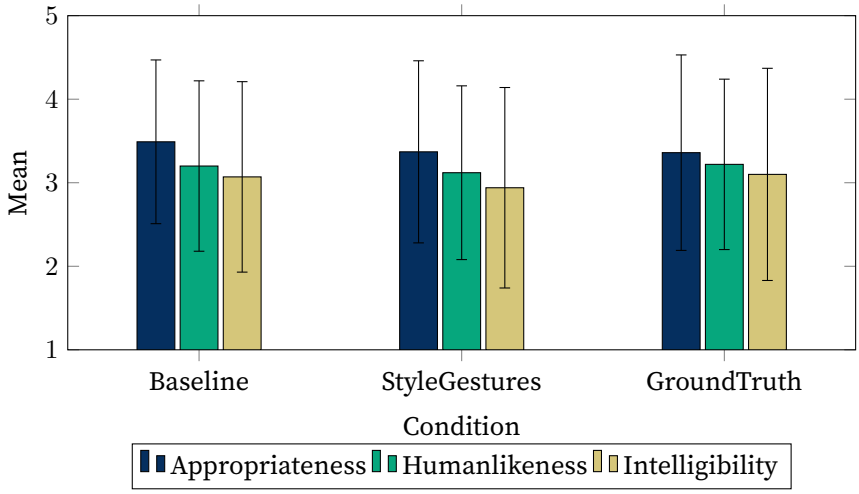


Figure 3.10 Mean and Error Bars for gesturing

degree of reliability among the items measuring intelligibility.

We performed a Mann Whitney U test for each construct between each condition (StyleGestures vs. Baseline, StyleGestures vs. Ground-truth, Baseline vs. Ground-truth). There were no significant differences between the scores on each construct. The mean scores are visualised in figure 3.10.

Listening

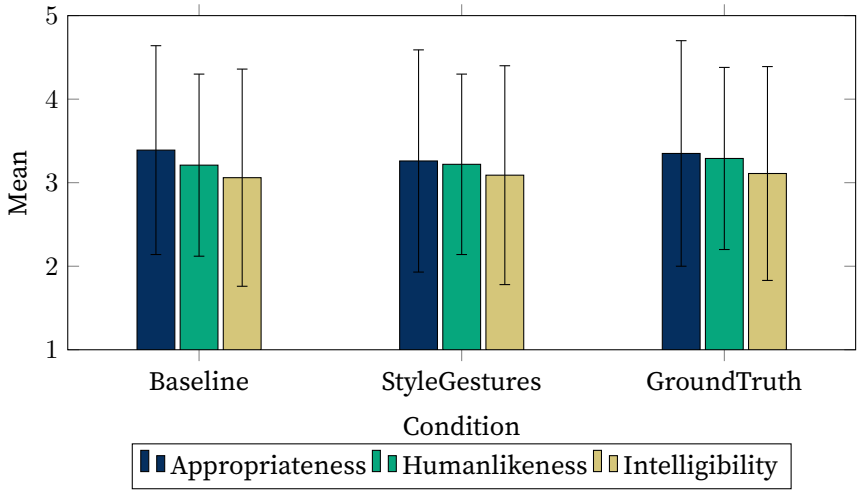


Figure 3.11 Mean and Error Bars for Each Condition in study 6.

We examined the perceived quality of synthesised listening motion across three dimensions: appropriateness, human-likeness, and intelligibility. To assess the internal consistency of the rating scales used for these dimensions, we calculated Cronbach's alpha coefficients. 48 participants (27 males, 21 females) participated and passed the attention checks. Of these participants, the mean age was 37 (SD=13). 39 participants resided in the UK, 3 in Ireland, 3 in Canada and 3 in Australia.

For the appropriateness dimension, the Cronbach's alpha coefficient was found to be 0.90 (95% CI [0.88, 0.91]), indicating good internal consistency and agreement among the items assessing appropriateness. For the human-likeness Cronbach's alpha coefficient was 0.93 (95% CI [0.92, 0.94]). For the intelligibility construct, Cronbach's alpha coefficient was found to be 0.98 (95% CI [0.98, 0.98]).

We performed a Mann Whitney U test for each construct between each condition (StyleGestures vs. Baseline, StyleGestures vs. Ground-truth, Baseline vs. Ground-truth). There were no significant differences between the scores on each construct. The mean scores are visualised in figure 3.11.

3.4 Synthesis and Discussion

In this section we discuss the three different studies we executed that looked into several evaluation methodologies. First, we discuss each section individually, followed by a global synthesis and conclusion.

3.4.1 Beat Gesture Generation: Model vs. Handcrafted Gestures

We can observe the clear preference for machine generated beat gestures, followed by designed beat gestures. The ranking of the last three conditions is well aligned with the amount of beat gestures in each condition; since condition two has the most gestures, followed by condition three and condition four. Using a ranking approach for evaluating generated gestures is not extremely informative. We got an idea of preferred conditions, but this only tells us exactly that, the preferences of one approach over the other. Using rankings, we were not able to deduce the quality of the gestures or the appropriateness of the gestures for the speech, or whether including gestures at all improved comprehension.

3.4.2 Rating vs. Comparison Evaluation Methods: Implications for Co-Speech Gesture Generation

We explored the differences in evaluating gesture motion stimuli with both pairwise comparisons and rating scales. Our aim was to gain a deeper un-

derstanding of when to use each approach. For this, we looked at the stimulus rankings both methods provided, their inter-rater reliability, the time it took participants to complete evaluations, participant preferences, and the conclusions both methods would yield regarding the comparison of gesture generation methods with high and low differences in quality.

The rank-ordering of stimuli between the pairwise comparisons and rating scales had a moderate positive correlation. We can conclude that in order to rank stimuli, in this instance co-speech gestures, there is not one approach that is preferred over the other; both are able to subjectively distinguish bad from good stimuli and this can be used to establish an order of quality.

When we take a look at the inter-rater reliability, we see a higher reliability for the pairwise method. This suggests that the pairwise method might be preferred over the rating scale method in terms of reliability.

When we look at which approach is faster, we can conclude that each comparison using the pairwise method was, on average, 6 seconds faster (25s instead of 31s) than each comparison using the rating scale method, which aligns with the findings of previous studies (Clark et al., 2018). Although this difference was statistically significant (i.e., reliable), a difference of 6 seconds per comparison is likely too small to make much of a practical difference unless the number of comparisons being made by each participant was large (e.g., 100 or more).

Whether participants reliably preferred one comparison method over the other depended on which method they were assigned to use first. Those participants who used the rating scale method and then the pairwise method significantly preferred the pairwise method. However, those who used the pairwise method and then the rating scale method did not show a reliable preference for either method. This provides tentative evidence that the pairwise method may be more user-friendly.

In line with a previous study (Kucherenko et al., 2020), we found that a high qualitative difference is indeed picked up by subjective evaluations. Not only does this hold for pairwise comparisons, but also for the rating scale approach. Both methods can provide similar results and are equivalent when comparing two or more conditions, for example two different models used to generate behaviour.

We did not find strong evidence to prefer one evaluation method over the other. However, the study we presented in this section does provide us some pointers and recommendations we can make in relation to the outcomes from this section:

Pairwise comparisons may be better suited when a large number of stimuli are to be evaluated, as this not only results in a shorter study but is likely to avoid fatigue in participants. If only a small number of conditions are under consideration, then pairwise comparisons of conditions is practical, but as the number of combinations grows with the faculty of the number of conditions ($\frac{n!}{2(n-2)!}$, with n the number of conditions) pairwise comparisons

tend to become unwieldy for 4 or more conditions if we want to compare all versus all.

Rating scales may be more appropriate when fine-grained evaluations are needed, as ratings can not only be used to rate stimuli between conditions, but can also be used to rank stimuli within conditions. Ratings are also recommended when more than 3 conditions are under considerations, as the number of required ratings grows linearly with the number of conditions and stimuli. We would however like to emphasise the importance of providing anchors/labels for each response option in the rating scales (Weijters et al., 2010). When using rating scales, it is also recommended to calibrate participants' judgements by showing the participants poor and excellent stimuli during a brief training session. While the lack of calibration can somewhat be addressed by normalising participants' ratings, resolution and reliability are lost when participants are not properly trained before starting their rating task.

Finally, it is important to consider the type of information provided by each evaluation method. Rating scales provide information about the quality of each stimulus on an *absolute* scale, whereas pairwise comparisons provide information on a *relative* scale. Thus, you could use the pairwise comparison method to establish whether one method of generating human-like behaviour was reliably preferred over another. However, being 'better' is not always the same as being 'good'. For instance, one method could be considered 'poor' and the other 'very poor'; this would likely result in a big difference in pairwise comparisons, but it would be a mistake to conclude that the former was therefore high quality in absolute terms. This is where carefully crafted rating scales (and qualitative methods, such as interviews and free response boxes) can provide additional information about quality in general.

3.4.3 Questionnaire Creation and Evaluation: Implications for Co-Speech Gesture Generation

We designed a questionnaire with appropriateness, human-likeness and intelligibility as constructs. We based our choice of statements and constructs on earlier work, reported by Wolfert, Robinson, and Belpaeme (2022). The internal consistency, measured through cronbach's Alpha, is high. For each construct for both studies, cronbach's Alpha is equal or higher than 0.9. This provides an indication that the statements together measure the intended construct. However, there were no significant differences in scores between the systems, not for gesticulation or listening behaviour evaluations. We can see small differences when we look at the figures, but these differences are not statistically significant. The evaluation of nonverbal behaviour poses a challenging task, as evidenced by the multitude of diverse evaluation paradigms employed over the years (Wolfert, Robinson, & Belpaeme, 2022). Unfortunately, there is a lack of a standardised and unified approach to measuring nonverbal behaviour, further complicating the evaluation process. In work by

he et al. (2022), multiple ways of measuring were applied to test the gesticulation of an avatar in an interaction, and only the behavioural method (through gaze tracking) yielded significant differences. Here, also a construct from the Godspeed (Bartneck et al., 2009) was included. Because of the high internal consistency, one could argue that multiple statements were measuring the same thing, reducing the resolution of the questionnaire, or that the Godspeed questionnaire is not a good questionnaire for evaluating human-like motion. On top of that, we only provided 5 answer options, which is a common way of applying Likert rating scales (Schrum et al., 2020). A possible explanation of the non-significant results is that the 5 answer options are not enough, and it is not sufficient for picking up these small differences in generated nonverbal behaviour. The number of statements participants had to answer per video could also have led to fatigue in participants, even with the low number of videos presented to the participant.

3.4.4 Summary of Findings and contributions

This chapter reported on three studies that covered four different evaluation approaches. In the first section, a ranking approach is used to decide on the best approach of generating beat gestures. Although we got a clear winner for the ranking, this approach does not seem to provide that much information. Following that, we looked at pairwise comparisons versus rating individual stimuli. These approaches can also be used to deduce a ranking, or to directly compare different systems. The final study that we reported on concerned the design and evaluation of using a questionnaire. Here we found no significant differences between the three conditions that were used for validating the questionnaire.

The downside of using a direct approach (through rating or pairwise comparisons) is that you can only measure one construct. This is were we brought in the questionnaire, since we hoped that we could get more information when participants would individually evaluate stimuli and answer statements for each stimuli. However, in our evaluation we did not find significant differences between the three conditions for the three constructs on appropriateness, human-likeness and intelligibility.

Crowdsourcing Gesture Generation Systems and Evaluations

4

Crowdsourcing Gesture Generation Systems and Evaluations

4.1 Introduction

In the previous chapter we discussed how to compare and evaluate generated gestures. In this chapter we look at two challenges we organized that invited researchers to submit their gesture generation systems and evaluated the results with online (crowdsourced) user studies.

4.1.1 Motivation for Gesture Generation Challenges

As discussed in the background chapter, there are many different systems and approaches towards generating gestures applied to many different platforms such as virtual avatars and social robots. Besides there being many different systems, there is an equal variety of datasets available. These two factors make it hard to directly compare the different qualities of a new proposed system, and led to the idea of setting up a challenge for gesture generation. Such a challenge makes it possible to compare different outcomes, and to look at the gap between natural gesticulation and computer generated gesticulation. The idea behind these challenges, known as GENEA (which stands for "Generation and Evaluation of Non-verbal Behaviour for Embodied Agents"), is that everyone can participate, get access to the data, but also participates in the crowdsourced evaluation. Through disseminating the results after the challenge through presentations and papers, we help improving the field.

4.1.2 Overview of GENEA 2020 and GENEA 2022

The first edition of the GENEA Challenge took place in 2020, and sixteen teams signed up after a public call for participation. Five teams completed the challenge and submitted their stimuli for the evaluation study. The second edition of the GENEA Challenge took place in 2022, and had ten teams participating. For both challenges, data and code was provided. Data was cleaned and split over a training, validation and test set. As code, participating teams were

provided with two baseline systems, and code to visualize their results on a virtual avatar.

The challenges were concluded with an online crowdsourced evaluation, where participating systems were evaluated on both human-likeness and appropriateness. Participating teams were asked to write a paper with information about their system and to clarify their findings. Each challenge was then concluded with a workshop where results were presented, and where participating teams could present and discuss their work.

4.2 Challenge Task

4.2.1 Task Description

The task of participating teams was to submit data-driven systems that were capable of generating gestures. For the first edition, we posed the problem of speech-driven gesture generation in the following way: given input speech features s – which could involve either an audio waveform (a sequence of pressure samples) or text (a word sequence) or the combination of the two – the task is to generate a corresponding pose sequence $\hat{g}$ describing gesture motion an ECA might perform while speaking.

For the second challenge we made use of a dataset containing multiple dyadic conversations. We did not use the conversation as a whole, but selected parts of the dataset to only have a single speaker. In addition to the previously described task, we now also included the speaker ID. Teams could submit to either the upper-body tier, or the full-body tier. In the upper-body tier, the legs were fixed and the corresponding visualisation was focused on the upper-body.

4.2.2 Differences between GENEA 2020 and GENEA 2022

For the 2022 version, participating teams could submit to either the upper-body tier, the full-body tier, or both. We had several reasons for this 'split'. First, one could argue that human embodied conversation uses the full body for nonverbal communication. Given that full-body behaviour generation is computationally a harder problem (with more joints involved, and thus of a higher dimension), if this problem gets solved, restricted versions of the same problem can be assumed to be solved as well. However, one could also argue that most of the nonverbal communication happens in the upper-body, and that it is unclear how much of the lower-body contributes to co-speech gestures. Including the full body might as well harm the final evaluation and function as a distraction. Since it is not clear which perspective is the right one, both tiers were included in the 2022 challenge, whereas the earlier edition of the challenge only focussed on upper-body motion generation.

4.3 Challenge Data

4.3.1 Data used in 2020 and 2022

More information about the datasets that are discussed here can also be found in the background section. The data for the 2020 challenge was based on the Trinity Gesture Dataset (Ferstl & McDonnell, 2018), comprising 244 min of audio and motion-capture recordings of a male actor speaking freely on a variety of topics. We removed lower-body data, retaining 15 upper-body joints out of the original 69. Finger motion was also removed due to poor capture quality.

For the 2022 challenge, we wanted to expand the dataset to include finger motion, lower-body motion, and material from multiple speakers in dyadic interactions. We based our challenge on the Talking With Hands 16.2M gesture dataset (Lee et al., 2019), which comprises 50 hours of audio (captured by close-talking directional microphones) and motion-capture recordings of several pairs of people having a conversation freely on a variety of topics, recorded in distinct takes each about 10 minutes long.

At the time of the challenge in 2022, this was likely the largest dataset of parallel speech and 3D motion (in joint-angle space) publicly available in the English language. Parts of the dataset (46 out of 116 takes) were removed that lacked audio or had low motion-capture quality, especially for the fingers.

For both data sets, information from the speech was obtained by transcribing the audio recordings using Google Cloud automatic speech recognition (ASR), followed by a thorough manual review to correct recognition errors and add punctuation for both the training and test parts of the dataset. All names of non-fictive persons were removed and replaced by unique tokens in the transcriptions.

Participants were allowed to use external motion, but these were specified to only include the following databases:

- CMU Motion Capture Database [1]
- Motion Capture Database HDM05 [2]
- CMU Panoptic Studio dataset [3]

We restricted the list of external motion we allowed since other challenges found that system performance is often limited by the amount of training data that can be ingested, which is not something we were interested in.

Despite the fact that the 2022 dataset was dyadic by design, the 2022 challenge focused on generating one side of the conversation at a time, without awareness of the interaction partner. This is different from the 2020 dataset, that

[1] mocap.cs.cmu.edu/
[2] resources.mpi-inf.mpg.de/HDM05/
[3] domedb.perception.cs.cmu.edu/

contained a single actor talking freely about different topics, without having to interact with an interlocutor.

4.4 Teams and Systems

Sixteen teams signed up for the GENEA 2020 challenge. Five teams completed the challenge, and the other teams failed to submit their results in time. Two teams later clarified why they stepped out of the challenge. One team gave as reason that they had to deal with reduced manpower, not being able to make the deadline in time, whereas the other team was not satisfied with the results. None of the teams reported issues with the task or the data as the reason for stepping out.

For the 2022 challenge, a total of 10 teams participated in the evaluation, with 8 *entries* (a.k.a. *submissions*) to the full-body tier and 8 entries to the upper-body tier, meaning that some teams submitted for both tiers.

4.4.1 Systems

GENEA 2020

The final evaluation for the 2020 challenge covered 9 different *conditions* or *systems*. There were 2 baselines, based on work by both Kucherenko et al. (2020) and Yoon et al. (2019). 2 toplines, based on direct human motion recordings, and 5 submitted systems.

			Inputs used		Representation or features		Stochastic
Name or description	Origin	ID	Aud.	Text	Input speech	Motion	output?
Natural motion	-	N	✓	✓	–	–	✓
Mismatched motion	-	M	✗	✗	–	–	✓
Audio-only baseline	Kucherenko et al. Kucherenko et al., 2019	BA	✓	✗	MFCC	Exp. map	✗
Text-only baseline	Yoon et al. Yoon et al., 2019	BT	✗	✓	FastText†	Rot. matrix	✗
AlltheSmooth Lu et al., 2020	CSTR lab, UEDIN, Scotland	S...	✓	✗	MFCCs	Joint pos.	✗
Edinburgh CVGU Pang et al., 2020	CVGU lab, UEDIN, Scotland	S...	✓	✓	BERT† & mel-spectr.	Rot. matrix	✓
FineMotion Korzun et al., 2020	ABBYY lab, MIPT, Russia	S...	✓	✓	GloVe† & mel-spectr.	Exp. map	✗
Nectec Thangthai et al., 2020	HCCR unit, NECTEC, Thailand	S...	✓	✓	Phoneme, Spacy word vecs. †, MFCCs, & prosody	Exp. map	✗
StyleGestures Alexanderson, 2020	TMH division, KTH, Sweden	S...	✓	✗	mel-spectr.	Exp. map	✓

Table 4.1 Conditions participating in the GENEA 2020 evaluation. Teams are sorted alphabetically by name. The anonymised IDs of submitted entries begin with the letter 'S' followed by a second, randomly-assigned letter in the range A through E, but which letter is associated which each team is not revealed in order to preserve anonymity. † indicates a use of word vectors pretrained on external data.

Table 4.1 lists all conditions, together with participating team names and (abbreviated) affiliations. Following the practice established by the Blizzard Challenge, we anonymised the teams by not revealing which team was assigned which ID, but individual teams are free to disclose their ID if they wish. Papers from each team describing their submitted systems in detail are available in the proceedings of the GENEA Workshop 2020 [4]

The two toplines were:

N Natural motion capture from the actor for the input speech segment in question. Surpassing this system would essentially entail superhuman performance.

M *Mismatched* natural motion capture from the actor, corresponding to another speech segment than that played together with the video. This was accomplished by permuting the motion segments from condition N in such a way that no segments remained in its original position. This represents the performance attainable by a system that produces very human-like motion (same as N, so a topline), but whose behaviour is completely unrelated to the speech (and thus can be considered as a bottom line in terms of motion appropriateness for the speech).

Since there has been no previous general study that compares systems to each other and what the state of the art is, it is hard to identify the "best" baseline systems to use. Therefore the choice was more subjective and based on code availability, with the two baseline systems chosen from recent data-driven gesture-generation papers that had their code available and were easy to reproduce. These were:

BA The system from Kucherenko et al. (2019), which only takes speech audio into account when generating system output. This model uses a chain of two neural networks: one maps from speech to pose representation and another decodes representation to pose, generating motion frame by frame by sliding a window over the speech input.

BT The system from Yoon et al. (2019), which only takes text transcript information (which includes word timing information) into account when generating system output. This model consists of an encoder for text understanding and a decoder for frame-by-frame pose generation.

The baseline systems were updated to work well on the challenge material. For system BA, new hyperparameters were needed to provide satisfactory results. For system BT, the learning rate and loss weights were adjusted. The resulting motion in system BA was represented using the exponential map Grassia (1998), and smoothed using a Savitzky-Golay filter Savitzky and Golay (1964). For BT, pose representation were changed from 2D Cartesian coordinates into rotational matrices, 3 for each of the 15 joints.

[4] zenodo.org/communities/genea2020/

GENEA2022

The 2022 challenge included the same toplines and baselines. The reason for including the same baselines are two-fold: it makes it possible to compare between challenge years, and it makes it possible to track progress in the field (by having the comparisons to other systems). The two toplines were also included in the 2022 challenges.

Baseline or team name	Per-tier label		Inputs used			Hands	Techniques used					Frame-wise	Stoch. output	Smthd.
			Aud.	Text	Sp. ID	Fixed	AR	RNNs	SA	VAEs	Other			
GestureMaster	FSA	USQ	✓	✓	✓						Rules, MGs			✓
Forgerons	FSC	USO	✓			✓	✓	✓		✓		✓	✓	
DeepMotion	FSI	USJ	✓	✓		✓	✓		✓	✓	CNNs	✓	✓	
DSI	FSF		✓				✓	✓	✓					
UEA Digital Humans	FSG	USM	✓	✓	✓			✓				✓		
ReprGesture		USN	✓	✓	✓	✓	✓	✓	✓	✓	CNNs, GANs			✓
IVI Lab	FSH	USK	✓	✓	✓	✓	✓	✓				✓	✓	✓
FineMotion	FSD		✓	✓			✓	✓				✓		✓
Murple AI lab	Not revealed		✓				✓	✓			Norm. flows	✓	✓	
Text-only baseline	FBT	UBT		✓		✓	✓	✓				✓		✓
Audio-only baseline		UBA	✓					✓				✓		✓
TransGesture		USL	✓			✓	✓	✓				✓		✓

Table 4.2 Conditions participating in the GENEA 2022 evaluation. Conditions are ordered based on their median human-likeness scores from higher to lower.

Table 4.2 shows the teams and their respective systems.

Similarities and differences between submitted systems

When we look at the submitted systems, we see that all submitted systems from both years relied on the audio input. For both challenges, some also included the provided text transcriptions. Self attention techniques are more common in systems submitted to the GENEA 2022 challenge. Many more systems in the 2022 version rely on autoregression techniques. No system in the 2020 challenge made use of hand-crafted rules and motion matching. No other large differences are spotted between the techniques that are applied in both challenges.

4.5 Evaluation

This section covers the evaluation of the stimuli that were submitted for both challenges. In here, we discuss the different subjective and objective metrics we used for assessing the performance of participating teams.

Synthetic gesture motion was required to be submitted at 20 frames per second (fps) for GENEA 2020, or 30 frames per second for GENEA 2022, in BVH format.

For the 2020 edition, participating teams were asked to synthesise a motion sequence for 20 minutes of test speech. However, we provided participants with smaller test samples for the 2022 challenge. In both situations, participants were not allowed to post-process the data in any way.

4.5.1 Subjective Evaluation

Human-likeness

For both challenges, human-likeness was assessed using the HEMVIP (Human Evaluation of Multiple Videos in Parallel) methodology (Jonell et al., 2021). With HEMVIP, multiple motion examples are presented in parallel, and subjects to the user study are asked to provide a rating for each one. All stimulus videos on the same page (a.k.a. screen) of the evaluation corresponded to the same speech segment but different conditions. The advantage of this method, is that differences in rating between the different conditions can be analysed using pairwise statistical tests, which helps control for variation between different subjects and different input speech segments; see Jonell et al. (2021). The videos used in this type of evaluation had the audio removed, since it has been found that speech and gesture perception influence each other (Bosker & Peeters, 2021) and can confound motion evaluations (Jonell, Kucherenko,

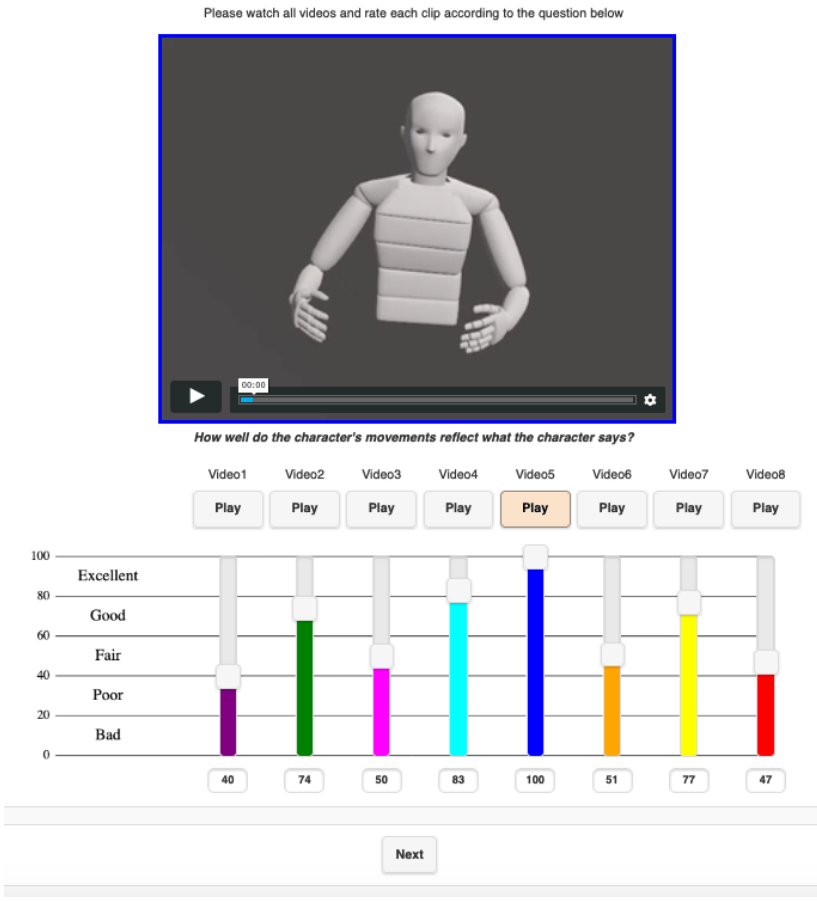


Figure 4.1 Screenshot of the rating interface from the study on appropriateness for GENEA 2020. The question asked in the image ("How well do the character's movements reflect what the character says?") originates from Jonell et al. (2021). This question was different for the human-likeness evaluation for both editions of the challenge.

Henter, & Beskow, 2020). Each HEMVIP page containing multiple videos, was accompanied by one question, where participants were asked "How human-like does the gesture motion appear?".

Figure 4.1 shows an example of the user interface used for the evaluation. The participants were first met with a screen with instructions and how to use the evaluation interface. For the 2022 evaluation, a training page was included directly after the instruction page. After that, they were then presented with 10 pages, where on each page they would compare and evaluate motion stimuli from all toplines, baselines, and most submitted systems, all for/with the same speech. It was possible for participants to return to previous conditions and change their rating after seeing other examples. Three attention checks were

incorporated into the pages for each study participant. These either displayed a brief text message over the gesticulating avatar reading “Attention! Please rate this video XX.”, or they temporarily replaced the audio with a synthetic voice speaking the same message. XX would be a number from 5 to 95, and the participant had to set the corresponding slider to the requested value, plus or minus 3, to pass the attention check. The numbers 13 through 19, as well as multiples of 10 from 30 to 90, were not used for attention checks due to their acoustic ambiguity. Which sliders on which pages that were used for attention check was uniformly random, except that no page had more than one attention check, and condition N and M were never replaced by attention checks. Lastly they were presented with a page asking for demographics and their experience of the test.

As can be seen in Figure 4.1, the 100-point rating scale was anchored by dividing it into successive 20-point intervals labelled (from best to worst) “Excellent”, “Good”, “Fair”, “Poor”, and “Bad”. These labels were based on those associated with the 5-point scale used for Mean Opinion Score (MOS) (International Telecommunication Union, Telecommunication Standardisation Sector, 1996) tests, another evaluation standard developed by the ITU.

Each study was balanced such that each segment appeared on pages 1 through 10 with approximately equal frequency across all raters (segment order), and each condition was associated with each slider with approximately equal frequency across all pages (condition order). For any given participant and study, each page would use different speech segments. Every page would contain condition N and (where relevant) condition M, but one other condition was randomly omitted from each page to limit the maximum number of sliders on a page to 8 or 7, depending on the study.

Appropriateness

For the first challenge, we decided to follow the setup from the human-likeness study, but with a different question. We asked participants “How appropriate are the gestures for the speech?”. For this task, stimuli included audio, as we aimed to investigate the link between motion and speech (both in terms of rhythm/timing and semantics), and ignoring motion quality.

For the second challenge we came up with a different design, following earlier work by Rebol et al. (2021). More specifically, we used the matching/mismatching paradigm. This takes two videos, one that matches with the audio, and one video that has no direct relation to the accompanying audio. This required a rework of the evaluation interface as well, of which a figure can be found in Chapter 3 on pairwise comparisons.

4.5.2 Stimuli

2020

For GENEA 2020, we selected 40 non-overlapping speech segments from the test inputs (average segment duration 10 s) to use in the user-study evaluation. These speech segments, which were not revealed to participants, were selected across the test inputs to be full and/or coherent phrases.

2022

For GENEA 2022, we selected 48 chunks from 40 test-set items, to be used in the subjective evaluations, since the dataset we used was dyadic by nature, we came up with a number of requirements for the test stimuli:

1. Segments should be around 8 to 10 seconds long, and ideally not shorter than 6 seconds.

2. The character should only be speaking, not passively listening, in the segments. (No turn-taking, but backchannels from the interlocutor were OK.)

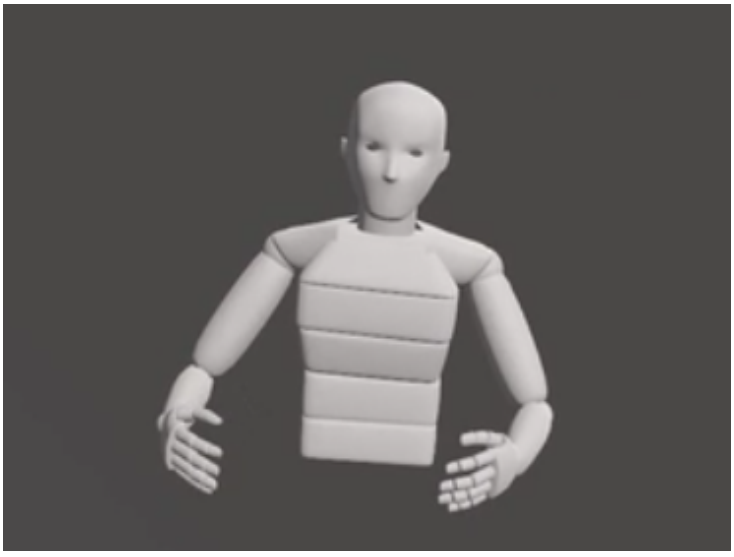
3. Segments should not contain any parts where Lee et al. (2019) had replaced the speech by silence for anonymisation.

4. Segments should be more or less complete phrases, starting at the start of a word and ending at the end of a word, and not end on a "cliffhanger". A small margin was permitted towards the end of segments.

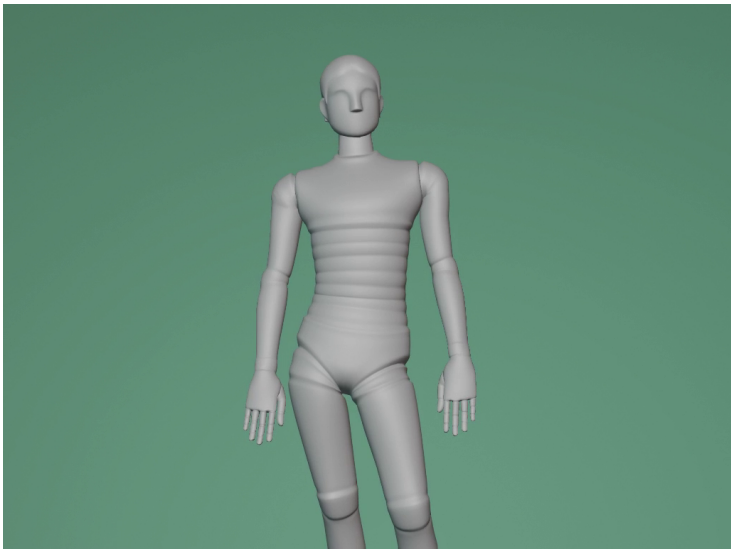
5. Finally, recorded motion capture in the segments (i.e., the FNA motion) should not contain any significant artefacts such as whole-body vibration or hands flicking open and closed due to poor finger tracking.

The last item does not imply that the motion capture was perfect or completely natural for all segments in the evaluation, since the finger-tracking quality throughout the database does not allow our evaluations to reach that standard. It merely means that the level of finger-tracking quality in the stimuli was consistent with the better parts of the source material from Lee et al. (2019).

The 48 selected segments were between 5.6 and 12.1 seconds in duration and on average 9.5 seconds long. Audio was loudness normalised to -23 dB LUFS following EBU R128 (Union, 2020) to maintain a consistent listening volume in the user studies.



(a) Avatar used in GENEA 2020



(b) Avatar used in GENEA 2022

Figure 4.2 Avatar comparison.

Visualisation

For GENEA 2020, the motion from the corresponding intervals in the BVH files submitted by participating teams was extracted and converted to a motion video clip using a visualisation server we provided to participants, at a resolution of 960×540 pixels.

We used the same virtual avatar for all renderings during the challenge and the evaluation. The avatar used in GENEA 2020 can be seen in Figure 4.2a. The avatar originally had 69 joints (full body including fingers) but only 15 joints, corresponding to the upper body and no fingers, were used for the challenge. Since hand and finger data had been omitted, these body parts were assigned a static pose, in which the hands were lightly cupped.

For GENEA 2022, a new avatar was designed in-house (see figure fig:comparison). This avatar had 56 joints (full-body including fingers). The avatar did not have eyes or a mouth, to make it easier for participants taking part in the evaluation to focus on the bodily movements. Participants were provided code for generating the visualisation, but they could also submit their BVH files for rendering videos, on a server provided by the organisers. The final motion was rendered in 30 FPS videos with a resolution of 1440×1000.

4.5.3 Test-participant recruitment

For both challenges, study participants were recruited through the crowdsourcing platform Prolific (formerly Prolific Academic), restricted to a set of English-speaking countries (UK, IE, USA, CAN, AUS, NZ). There was no requirement to be a native speaker of English, since Prolific does not support screening participants based on that criterion. A participant could take either study or both studies, but not more than once each.

For the GENEA 2020 user studies, participants were remunerated 5.75 GBP for completing the human-likeness study (median time 33 min) and 6.50 GBP for the appropriateness study (median time 34 min).

For the user studies part of GENEA 2022, participants were paid 6 GBP for each successfully completed human-likeness study (median time 28 min), and 5.5 GBP for the appropriateness study (median time 25 min). These compensation levels all exceed the UK national living wage.

4.5.4 Objective Metrics

The main goal of the GENEA challenges was to compare human subjective impressions of the outputs of different gesture-generation systems. This is the main reason we discouraged using the results of automated performance metrics as indicators for the perceptual quality of different systems. However, we calculated different objective metrics to identify possible objective

metrics that are meaningful and could have a relation to subjective metrics. If an objective metric would be meaningful, it could be used during system development.

For GENEA 2020 we took into account two objective metrics: average jerk and the distance between speed histograms. For the second edition, we extended the selection of objective metrics to a total of five measures, that have been used before in the field, namely average jerk, average acceleration, distance between gesture speed (i.e., absolute velocity) histograms, canonical correlation analysis, and the Fréchet distance between motion feature distributions.

4.6 Results

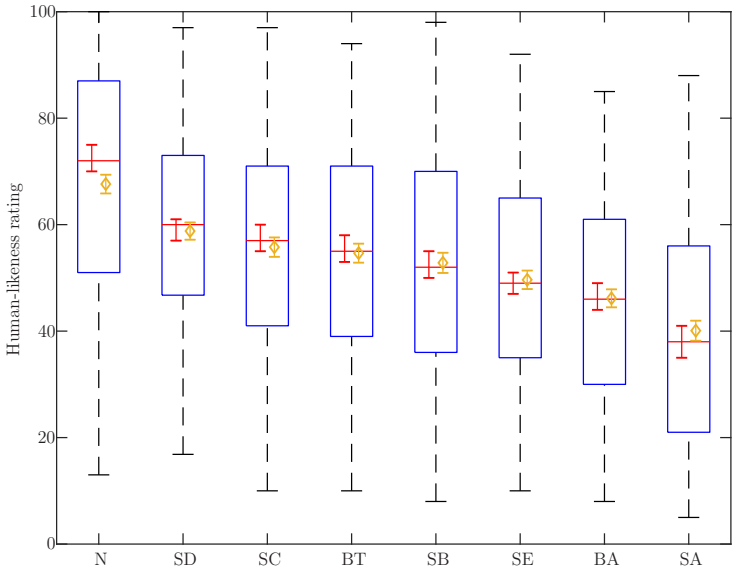
This section describes the results obtained from both GENEA challenges. First, we dive into the results we obtained during the first edition, followed by the results from the second edition.

4.6.1 GENEA 2020

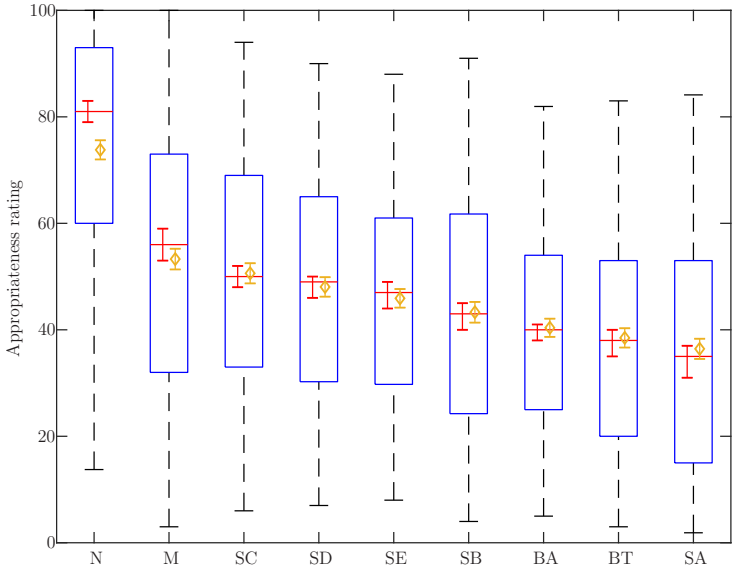
Data on test participants

For GENEA 2020, each user study recruited 125 participants. All participants passed the attention checks they encountered. In the human-likeness study, average reported participant age was 31.5 years (standard deviation 10.7), with 66 men, 57 women, and 2 others. We asked participants on which continent they lived, and 69 participants were from Europe, 1 from Africa, 48 from North America, 2 from South America, and 5 from Asia. In the appropriateness study, average age was 31.1 years (SD 11.7), with 60 men, 64 women, and 1 other. 78 participants reported residing in Europe, 1 in Africa, 39 in North America, 3 in Asia, and 4 in Oceania. Each study had 116 native and 9 non-native speakers of English.

23 test-takers in the human-likeness study and 40 test-takers in the appropriateness study did not pass all attention checks. These test-takers were not part of the 125 participants analysed. Scores from sliders used for attention checks were also omitted, leaving in total 8,375 and 9,625 ratings that were analysed in each of the two respective studies. The median successful completion time for the main part of the study was 24 min for the human-likeness study and 27 min for the appropriateness study, with the shortest successful completion time being 12 min in both studies. These figures exclude reading instructions and answering the post-test questionnaire.



(a) Human-likeness ratings



(b) Appropriateness ratings

Figure 4.3 Box plots visualising the ratings distribution in the two studies. Red bars are the median ratings (each with a 0.01 confidence interval); yellow diamonds are mean ratings (also with a 0.01 confidence interval). Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each system. Conditions are ordered descending by sample median, which leads to a different order in each of the two plots.

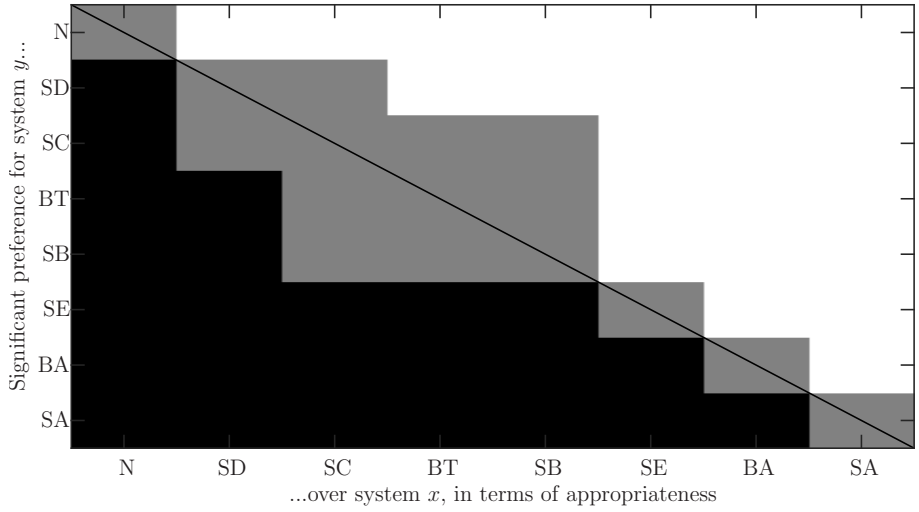
Analysis and results of subjective evaluation

Summary statistics (sample median and sample mean) for all conditions in each of the two studies are shown in Table 4.3 (see page 88), together with a 99% confidence interval for the true median/mean. The confidence intervals were computed either using a Gaussian assumption for the means (i.e., with Student's t-distribution cdf, and rounded outward to ensure sufficient coverage), or using order statistics for the median (leverages the binomial distribution cdf, cf. (Hahn & Meeker, 1991)).

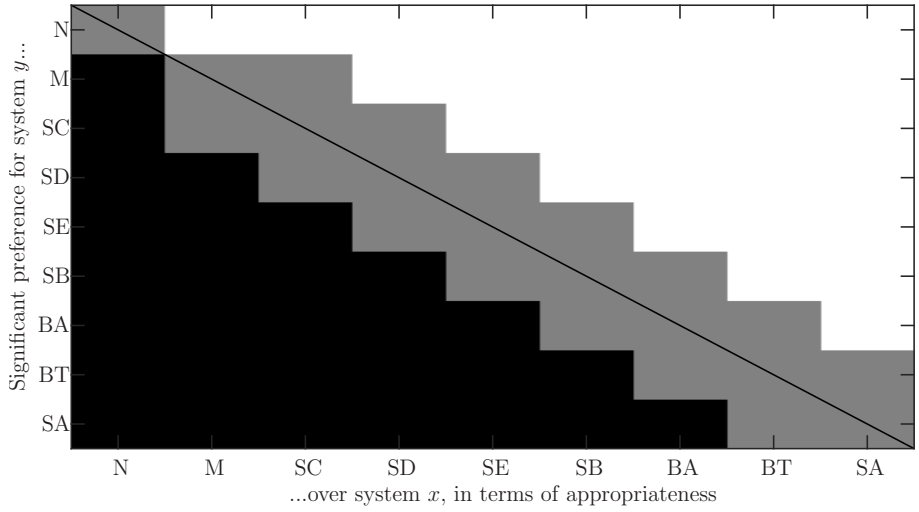
The ratings distributions in the two studies are further visualised through box plots in Figure 4.3. The distributions are seen to be quite broad. This is common in Multiple stimuli with hidden reference and anchor (MUSHRA)-like evaluations, since the range of numbers not only reflects differences between systems, but also extraneous variation, e.g., between stimuli, in individual preferences, and in how critical different raters are in their judgements. In contrast, the plotted confidence intervals are seen to be quite narrow, due to the large number of ratings collected for each condition.

Despite the wide range of the distributions, the fact that the conditions were rated in parallel on each page enables using pairwise statistical tests to factor out many of the above sources of variation. To analyse the significance of differences in sample median between different conditions, we applied two-sided pairwise Wilcoxon signed-rank tests to all pairs of distinct conditions in each study. This closely follows the analysis methodology used throughout recent Blizzard Challenges. (Unlike Student's t-test, this test does not assume that rating differences follow a Gaussian distribution, which would likely be inappropriate, as we can see from the box plots in Figure 4.3 that ratings distributions are skewed and thus non-Gaussian.) For each condition pair, only pages for which both conditions were assigned valid scores were included in the analysis. (Recall that not all systems were scored on all pages due to the limited number of sliders and the presence of attention checks.) This meant that every statistical significance test was based on at least 796 pairs of valid ratings in each of the studies. The p-values computed in the significance tests were adjusted for multiple comparisons using the Holm-Bonferroni method (Holm, 1979) (which is uniformly more powerful than regular Bonferroni correction) in each of the two studies. This statistical analysis found all but 4 out of 28 condition pairs to be significantly different in the human-likeness study, which the corresponding numbers being 7 out of 36 condition pairs in the appropriateness study, all at the level $\alpha = 0.01$. Which conditions that were found to be rated significantly above or below which other conditions in the two studies is visualised in Figure 4.4.

Finally, we present two diagrams that bring the results of the two perceptive studies for GENEA 2020 together. Figure 4.5, in particular, visualises the relative (partial) ordering between different conditions implied by the results of the two studies in Figure 4.4. Although there are similarities, the two orderings are meaningfully different. This, together with the results in (Jonell



(a) Human-likeness study



(b) Appropriateness study

Figure 4.4 Significance of pairwise differences between conditions. White means that the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at the 0.01 level after Holm-Bonferroni correction. Conditions are listed in the same order as in Figure 4.3, which is different for each of the two studies.

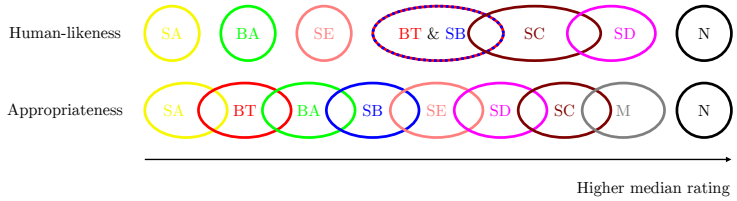


Figure 4.5 Partial ordering between conditions in the two studies. Each condition is an ellipse; overlapping or (in one case) coinciding ellipses signify that the corresponding conditions were not statistically significantly different in the evaluation. The diagram was inspired by Wester et al. (2016) with colours adapted from Boynton (1989). There is no scale on the axis since the figure visualises ordinal information only.

et al., 2021), reinforces a conclusion that the two studies managed to disentangle aspects of perceived motion quality (human-likeness) from the perceived link between gesture and speech (appropriateness). Figure 4.6, meanwhile, visualises confidence regions for the median rating as boxes whose horizontal and vertical extents are given by the corresponding confidence intervals in Table 4.3. Once again, different systems are found to be good at different things. The numerical gap between natural and synthetic gesture motion is seen to be more pronounced in the case of appropriateness than for human-likeness.

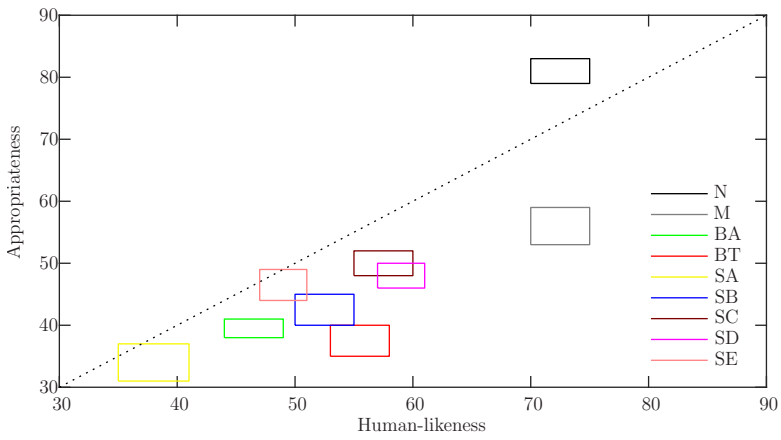


Figure 4.6 Confidence regions for the true median rating across both studies. The dotted black line is the identity, $x = y$. While the human-likeness (x-coordinate) of M was not evaluated directly, it is expected to be very close to N since it uses the same motion clips, and the horizontal extent of the confidence region for M was therefore copied from N.

Table 4.3 Summary statistics of user-study ratings for all conditions in the two studies, with 0.01-level confidence intervals. The human-likeness of M was not evaluated explicitly, since it uses the same motion clips as N.

ID	Human-likeness		Appropriateness	
	Median	Mean	Median	Mean
N	$72 \in [70, 75]$	67.6 ± 1.8	$81 \in [79, 83]$	73.8 ± 1.8
M	"	"	$56 \in [53, 59]$	53.3 ± 2.0
BA	$46 \in [44, 49]$	46.2 ± 1.7	$40 \in [38, 41]$	40.4 ± 1.8
BT	$55 \in [53, 58]$	54.6 ± 1.8	$38 \in [35, 40]$	38.5 ± 1.9
SA	$38 \in [35, 41]$	40.1 ± 1.9	$35 \in [31, 37]$	36.4 ± 1.9
SB	$52 \in [50, 55]$	52.8 ± 1.9	$43 \in [40, 45]$	43.3 ± 2.0
SC	$57 \in [55, 60]$	55.8 ± 1.9	$50 \in [48, 52]$	50.6 ± 1.9
SD	$60 \in [57, 61]$	58.8 ± 1.7	$49 \in [46, 50]$	48.1 ± 1.9
SE	$49 \in [47, 51]$	49.6 ± 1.8	$47 \in [44, 49]$	45.9 ± 1.8

Table 4.4 Results from the objective evaluations. The Hellinger distance between natural and synthetic speed profiles was computed for the two wrist joints, since hand motion is of central importance for co-speech gestures.

ID	Jerk	Hellinger distance	
		Left	Right
N	151.52 ± 35.57	0	0
BA	65.59 ± 4.42	0.084	0.090
BT	45.84 ± 2.14	0.130	0.096
SA	132.37 ± 27.64	0.064	0.059
SB	189.39 ± 4.66	0.126	0.114
SC	84.44 ± 8.48	0.083	0.088
SD	72.06 ± 7.91	0.073	0.062
SE	97.85 ± 9.34	0.049	0.049

Results of objective evaluation

Results of the objective evaluations from Section 4.5.4 are given in Table 4.4. The first column contains the average jerk across all the joints. We report mean and standard deviation for the full 20 min of test motion. The second and third columns contain the Hellinger distance between speed histograms for the left and right wrists.

Different systems performed best (coming closest to the natural motion N) in different objective measures. For example, systems SA and SB where the closest to the ground truth in terms of the jerk value, but SE and SD were among the closest to the ground truth as measured by Hellinger distance between speed histograms.

We also found that objective metrics deviate from the subjective results. While SA showed the most similar jerk to natural motion, it was less preferred in the subjective evaluation. Similarly, SE showed the Hellinger distances most similar to N, but was not close to being the most preferred synthetic system in the subjective evaluation. Considering this disparity, we stress that objective evaluation of gesture motion is a complementary measure, and that subjective evaluation is much more important.

4.6.2 GENEA 2022

Data on test participants

For the human-likeness study, we recruited 121 test takers that successfully passed the attention checks and completed the full-body study, and 150 test takers that successfully passed the attention checks and completed the upper-body study. Of the 121 test takers in the full-body study, 60 identified as female, 60 as male, and 1 did not want to disclose their gender. The same numbers for the 150 upper-body test takers were 74, 75, and 1, respectively. For the full-body test takers, 2 resided in Australia, 2 in Canada, 3 in Ireland, 110 in the United Kingdom, and 4 in the United States of America. The upper-body study had 1 Australian resident, 4 from Ireland, 134 from the United Kingdom, and 11 from the United States of America.

For the appropriateness studies, our design goal was for each condition to receive as many responses per condition as the number of ratings that each condition (aside from FNA/UNA) received in the corresponding human-likeness evaluation. This works out to 880 responses per condition in the full-body studies and 990 responses per condition in the upper-body studies. Because a subject in these studies provided half as many responses as in a human-likeness study (40 vs. 80), the appropriateness studies needed to recruit approximately twice as many test takers.

In the end, 247 test takers successfully passed the attention checks in the full-body study, while 304 passed the attention checks in the upper-body study. Of

the 247 subjects in the full-body study, 137 identified as female, 107 as male, and 3 did not want to disclose their gender. The same numbers for the 304 upper-body test takers were 127, 173, and 4, respectively. For the full-body test takers, 3 resided in Australia, 13 in Canada, 10 in Ireland, 2 in New Zealand, 211 in the United Kingdom, and 8 in the United States of America. The upper-body study had 2 residents from Australia, 10 from Canada, 1 from Ireland, 256 from the United Kingdom, and 35 from the United States of America.

Analysis and results of subjective evaluation: human-likeness

Each test taker in the human-likeness studies contributed 76 ratings to the analyses after removing attention checks, giving a total of 9,196 ratings for the full-body study and 11,400 ratings for the upper-body study. The results are visualised in Figure 4.7, with summary statistics for the ratings of all conditions given in the first half of Table 4.5, together with 95% confidence intervals for the true median. These confidence intervals were computed using order statistics, leveraging the binomial distribution cdf; see Hahn and Meeker (1991).

Table 4.5 Summary statistics of responses from all user studies of GENEA 2022, with 95% confidence intervals. "M." stands for "matched" and "Mism." for "mismatched". "Percent matched" identifies how often subjects preferred matched over mismatched motion.

(a) Full-body

ID	Median human-likeness	Appropriateness Num. responses M.	Tie	Mism.	Percent matched (splitting ties)
FNA	$70 \in [69, 71]$	590	138	163	$74.0 \in [70.9, 76.9]$
FBT	$27.5 \in [25, 30]$	278	362	250	$51.6 \in [48.2, 55.0]$
FSA	$71 \in [70, 73]$	393	216	269	$57.1 \in [53.7, 60.4]$
FSB	$30 \in [28, 31]$	397	163	330	$53.8 \in [50.4, 57.1]$
FSC	$53 \in [51, 55]$	347	237	295	$53.0 \in [49.5, 56.3]$
FSD	$34 \in [32, 36]$	329	256	302	$51.5 \in [48.1, 54.9]$
FSF	$38 \in [35, 40]$	388	130	359	$51.7 \in [48.2, 55.1]$
FSG	$38 \in [35, 40]$	406	184	319	$54.8 \in [51.4, 58.1]$
FSH	$36 \in [33, 38]$	445	166	262	$60.5 \in [57.1, 63.8]$
FSI	$46 \in [45, 48]$	403	178	312	$55.1 \in [51.7, 58.4]$

(b) Upper-body

ID	Median human-likeness	Appropriateness Num. responses M.	Tie	Mism.	Percent matched (splitting ties)
UNA	$63 \in [61, 65]$	691	107	189	$75.4 \in [72.5, 78.1]$
UBA	$33 \in [31, 34]$	424	264	303	$56.1 \in [52.9, 59.3]$
UBT	$36 \in [34, 39]$	341	367	287	$52.7 \in [49.5, 55.9]$
USJ	$53 \in [52, 55]$	461	164	365	$54.8 \in [51.6, 58.0]$
USK	$41 \in [40, 44]$	454	185	353	$55.1 \in [51.9, 58.3]$
USL	$22 \in [20, 25]$	282	548	159	$56.2 \in [53.0, 59.4]$
USM	$41 \in [40, 42]$	503	175	328	$58.7 \in [55.5, 61.8]$
USN	$44 \in [41, 45]$	443	190	352	$54.6 \in [51.4, 57.8]$
USO	$48 \in [47, 50]$	439	209	335	$55.3 \in [52.1, 58.5]$
USP	$29.5 \in [28, 31]$	440	180	376	$53.2 \in [50.0, 56.4]$
USQ	$69 \in [68, 70]$	504	182	310	$59.7 \in [56.6, 62.9]$

To analyse the significance of differences in median rating between different conditions, we applied two-sided pairwise Wilcoxon signed-rank tests to all unordered pairs of distinct conditions in each study. For each condition pair, only cases where both conditions appeared on the same page were included in the analysis of significant differences. Because this analysis is based on pairwise statistical tests, it can potentially resolve differences between conditions that are smaller than the width of the confidence intervals for the median in Figure 4.7, since those confidence intervals are inflated by variation that the statistical test controls for. The p-values computed in the significance tests were adjusted for multiple comparisons on a per-study basis using the Holm-Bonferroni method (Holm, 1979).

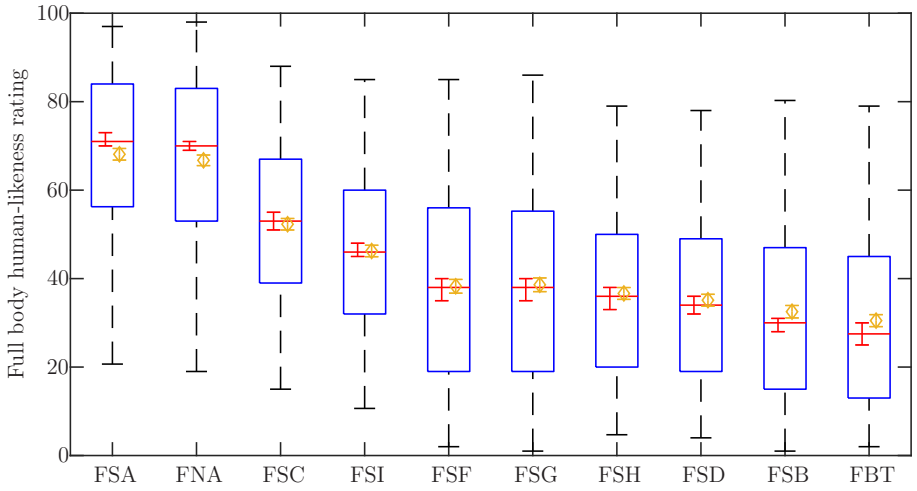
Our statistical analysis found all but 5 out of 45 condition pairs to be significantly different in the full-body study and all but 2 out of 55 condition pairs to be significantly different in the upper-body study, all at the level $\alpha = 0.05$ after Holm-Bonferroni correction. The significant differences we identified are visualised in Figure 4.8.

Analysis and results of subjective evaluation: appropriateness

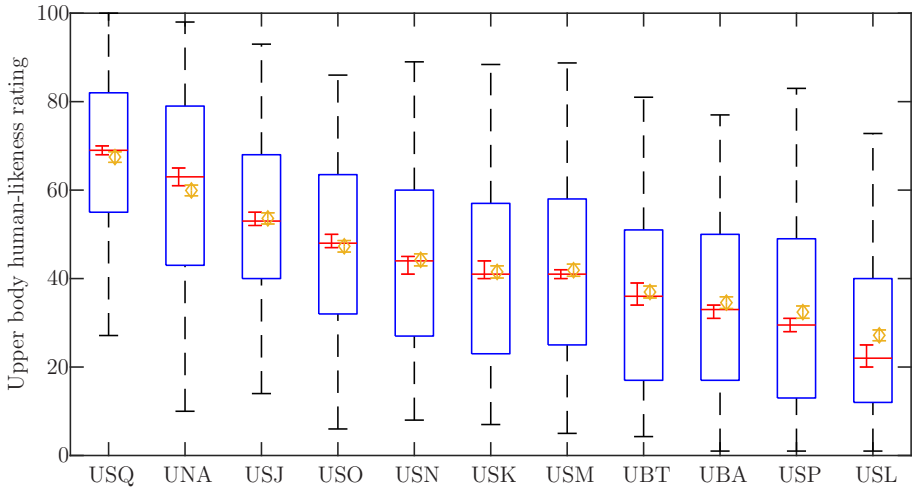
We gathered a total of 8,867 responses for the full-body study and 10,910 responses from the upper-body study that were included in the analysis.

Raw response statistics for all conditions in each of the two studies are shown in the second half of Table 4.5, together with 95% Clopper-Pearson confidence intervals for the fraction of time that the matched video was preferred over the mismatched, after dividing ties equally between the two groups (rounding up in case of non-integer counts). The quoted confidence intervals were rounded outward to ensure sufficient coverage.

The response distributions in the two studies are further visualised through bar plots in Figure 4.9, while Figure 4.10 visualises the results of the entire challenge in a single coordinate system per tier. Overall, the distribution of the three different responses across the different conditions is consistent with the mismatching study reported in Jonell, Kucherenko, Henter, and Beskow, 2020. No system has a relative preference for matched motion below 50%, which is the theoretical bottom line, attained by a system whose motion has no relation to the speech. (Here and forthwith, we only consider the relative preference in the sample after dividing ties equally.) The greatest relative preference, a 75% preference for matched motion, is observed for natural motion capture, i.e., FNA/UNA. This should be considered a good result, since previous studies that have incorporated mismatched stimuli, e.g., Jonell, Kucherenko, Henter, and Beskow (2020) and Kucherenko, Jonell, Yoon, Wolfert, and Henter (2021), have found that they sometimes are difficult for participants to distinguish from matched ones, especially if they – like here – both correspond to segments where the character is speaking. Furthermore, both matched and mismatched motion stimuli have their starting points aligned to the start of a phrase in the speech, meaning that the motion

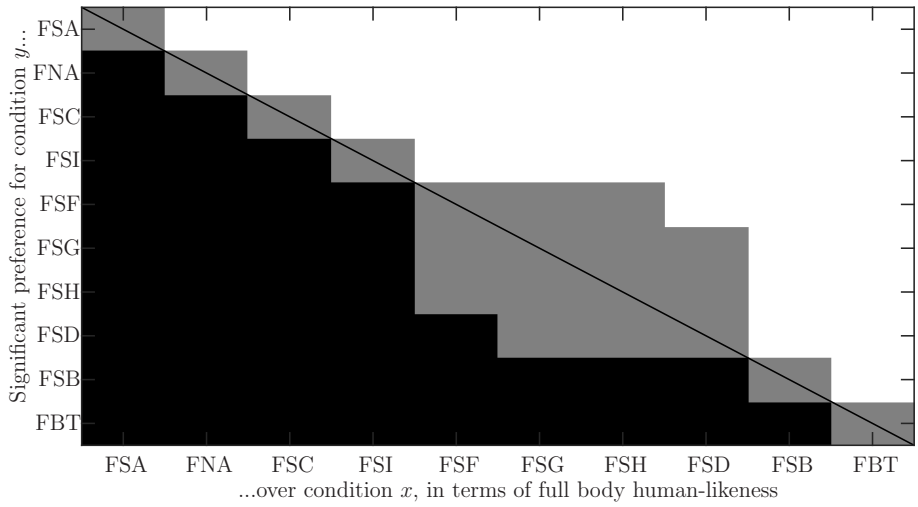


(a) Full-body

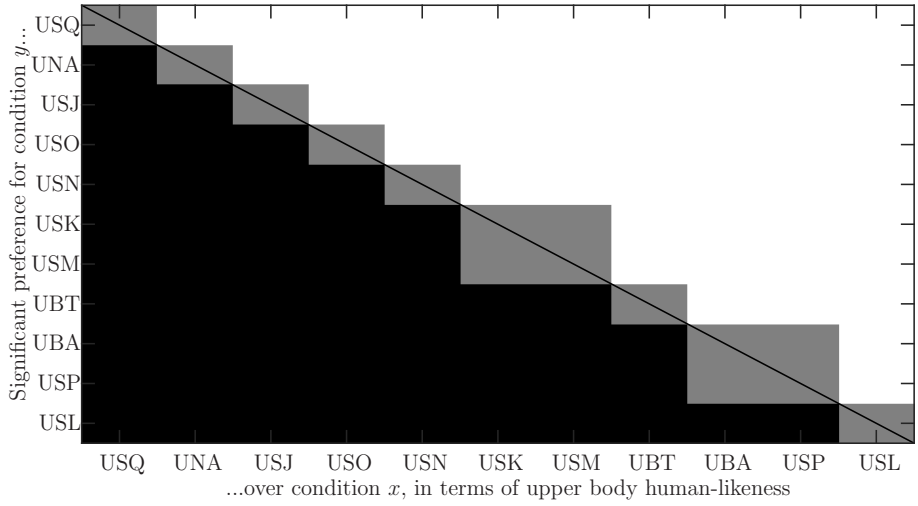


(b) Upper-body

Figure 4.7 Box plots visualising the ratings distribution in the human-likeness studies. Red bars are medians and yellow diamonds are means, each with a 0.05 confidence interval and a Gaussian assumption for the means. Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each condition. Conditions are ordered descending by sample median for each tier.

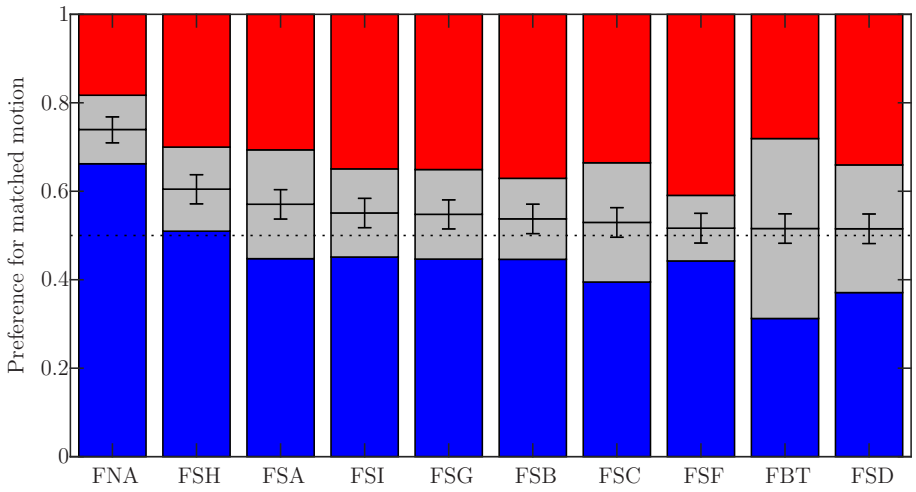


(a) Full-body

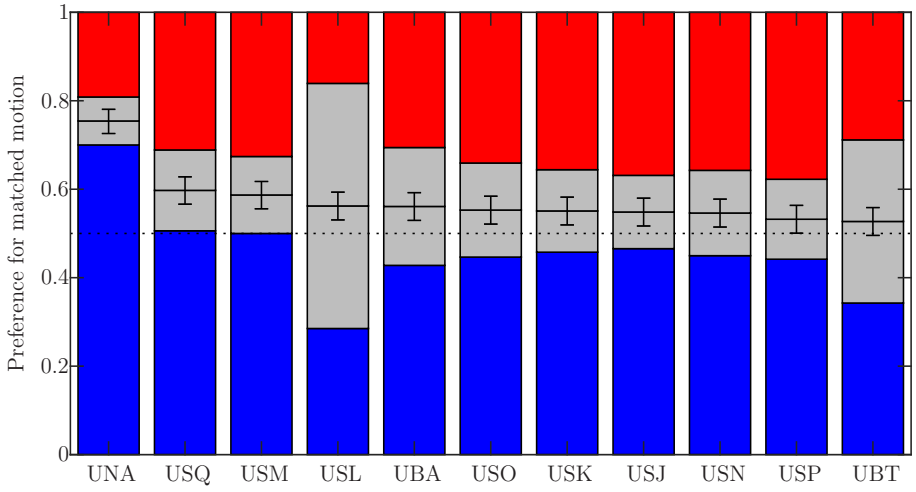


(b) Upper-body

Figure 4.8 Significant differences in human-likeness. White means the condition listed on the y-axis rated significantly above the condition on the x-axis, black means the opposite (y rated below x), and grey means no statistically significant difference at level $\alpha = 0.05$ after Holm-Bonferroni correction. Conditions use the same order as the corresponding subfigure in Figure 4.7.

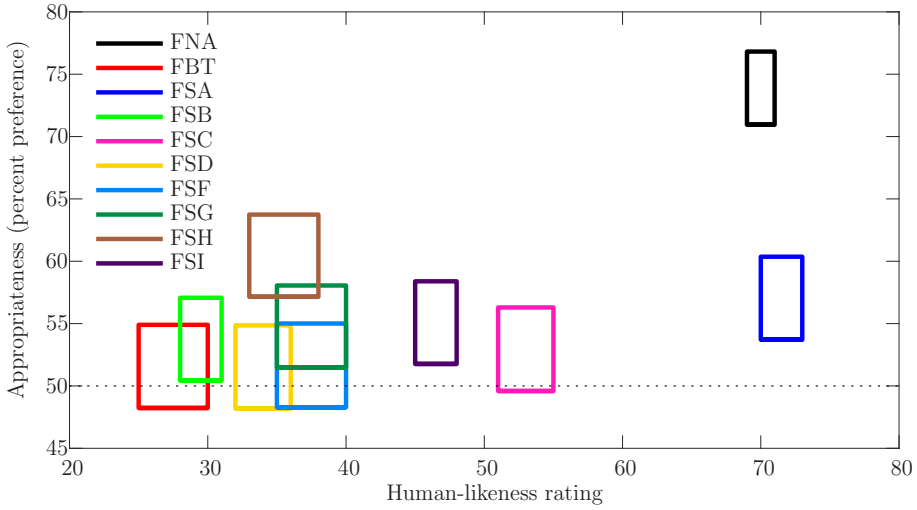


(a) Full-body

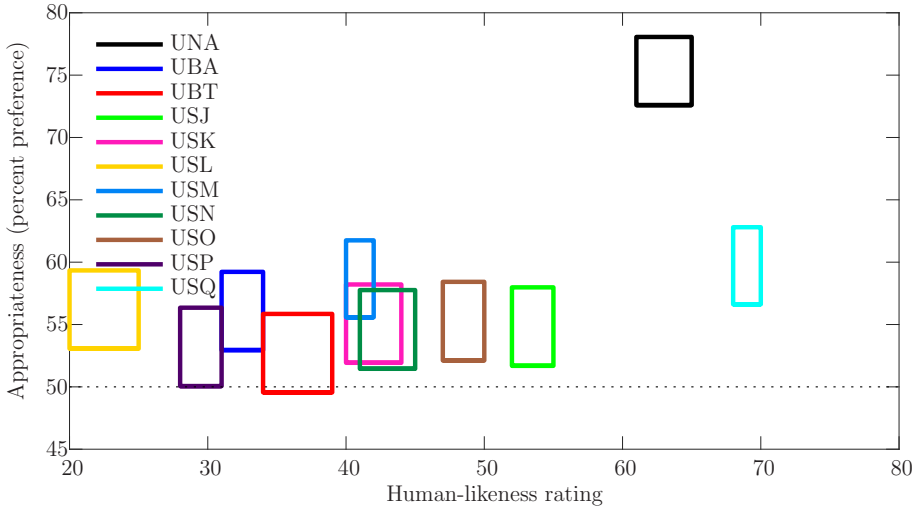


(b) Upper-body

Figure 4.9 Bar plots visualising the response distribution in the appropriateness studies. The blue bar (bottom) represents responses where subjects preferred the matched motion, the light grey bar (middle) represents tied ("They are equal") responses, and the red bar (top) represents responses preferring mismatched motion, with the height of each bar being proportional to the fraction of responses in each category. The black horizontal lines bisecting the light grey bars represent the proportion of matched responses after splitting ties, each with a 0.05 confidence interval. The dotted black line indicates chance-level performance. Conditions are ordered by descending preference for matched motion after splitting ties.



(a) Full-body



(b) Upper-body

Figure 4.10 Joint visualisation of the evaluation results for each tier. Box widths show 95% confidence intervals for the median human-likeness rating and box heights show 95% confidence intervals for the preference for matched motion in percent, indicating appropriateness. FNA and UNA are the natural recorded motion.

in the stimulus videos might initially be more similar to each other than if the mismatched motion had been excerpted completely at random and not aligned to the start of phrase boundaries.

Unlike the human-likeness studies, the responses in the appropriateness studies are restricted to three categories and do not necessarily come in pairs for statistical testing in the same way as for the parallel sliders in HEMVIP. A different method for identifying significant differences therefore needs to be adopted. We used Barnard's test (Barnard, 1945) to identify statistically significant differences at the level $\alpha = 0.05$ between all pairs of distinct conditions, applying the Holm-Bonferroni method (Holm, 1979) to correct for multiple comparisons as before. Barnard's test is considered more appropriate than Fisher's exact test for a product of two independent binomial distributions (Lydersen et al., 2009), as here. This analysis found 13 of 45 condition pairs to be significantly different in the full-body study and 10 out of 55 condition pairs to be significantly different in the upper-body study. Specifically, FNA/UNA were significantly more appropriate for the specific speech signal compared to all other, synthetic conditions. In addition, FSH was significantly more appropriate than FBT, FSC, FSD, and FSF in the full-body study. No other differences were statistically significant in either study.

Instead of comparing the appropriateness of different synthesis approaches against one another, one can compare against a random baseline (50/50 performance), and test if the observed effect size is statistically significantly different from zero. We can assess this at the 0.05 level by checking whether or not the confidence interval on the effect size overlaps with chance performance. From this perspective, FSA, FSB, FSG, FSH, FSI are significantly more appropriate than chance in the full-body study, and all systems except UBT are more appropriate than chance in the upper-body study. Unlike other significance tests in this chapter, these do not include a correction for multiple comparisons.

Results of objective evaluation

The values of the objective metrics we computed are listed in Table 4.6. For each number in the table, we also calculated how much it differed from the corresponding value for the reference system (FNA/UNA), and then computed the rank correlation between the absolute value of these differences and the median human-likeness scores from the subjective evaluation. The idea is that systems exhibiting values closer to FNA/UNA should appear more human-like.

The resulting rank correlations and p-values can be found in Table 4.7. For median human-likeness, we only found a statistically significant ($p < 0.05$) rank correlation with FGD (Fréchet Gesture Distance), for both the full and upper-body tier (Kendall's $\tau = -0.49$ and -0.51, respectively). The negative sign is expected, since a smaller difference from FNA/UNA should be associated with better-looking motion and higher human-likeness scores. Fig. 4.11 visually compares the subjective human-likeness ratings and objective metric results.

CCA is the only metric we computed that can indicate appropriateness, since it directly compares each generated sequence to the corresponding reference motion-capture poses. We therefore computed its rank correlations with the appropriateness data as well. Here we found a statistically significant effect ($\tau = -0.49$) for the upper-body tier, but not for the full body.

(a) Full-body

Condition	Average jerk	Average accel.	Global CCA	Hellinger distance	FGD
FNA	**31300 ± 6590**	**798 ± 208**	**1**	**0**	**0**
FSA	14600 ± 2970	**668 ± 161**	0.849	**0.041**	**3.18**
FSC	5130 ± 2120	332 ± 129	0.818	0.125	16.4
FSI	7370 ± 1710	345 ± 98	0.789	0.111	**4.87**
FSF	**22600 ± 6240**	**666 ± 223**	**0.916**	0.195	7.49
FSG	5560 ± 2380	282 ± 127	**0.992**	**0.060**	10.1
FSH	8630 ± 2440	313 ± 92	**0.968**	0.104	**4.02**
FSD	8690 ± 8320	405 ± 257	0.886	0.132	43.4
FSB	**27200 ± 4680**	**628 ± 116**	0.782	**0.050**	16.3
FBT	3510 ± 1090	177 ± 56	0.738	0.267	28.6

(b) Upper-body

Condition	Average jerk	Average accel.	Global CCA	Hellinger distance	FGD
UNA	**33000 ± 7030**	**842 ± 222**	**1**	**0**	**0**
USQ	**15400 ± 3190**	**710 ± 173**	0.685	**0.043**	**2.84**
USJ	8280 ± 1460	375 ± 81	0.640	0.197	**4.83**
USO	5450 ± 2260	353 ± 138	0.812	0.129	16.4
USN	7510 ± 3400	384 ± 127	0.789	0.092	194
USK	8180 ± 2450	311 ± 99	**0.962**	0.137	15.5
USM	6840 ± 3200	385 ± 172	**0.991**	**0.039**	**2.17**
UBT	3760 ± 1170	190 ± 60	0.707	0.248	18.2
UBA	**18000 ± 14900**	513 ± 326	**0.964**	0.244	17.0
USP	**28500 ± 4960**	**661 ± 123**	0.769	**0.051**	18.0
USL	7730 ± 5420	258 ± 157	0.849	0.306	28.4

Table 4.6 Objective evaluation results for the GENEA 2022 Challenge. The word "acceleration" has been abbreviated to "accel."; ± shows the standard deviation per sequence. The best two or three numbers in each column, i.e., those closest to the numbers from the held-out motion-capture data (FNA/UNA, first row of values), are bold. Except for FNA/UNA, conditions (rows) are ordered by decreasing median human-likeness rating. Numbers have generally been rounded to three significant digits.

(a) Full-body

Metric	Average jerk	Average accel.	Global CCA		Hellinger distance	FGD
Versus	Hum.	Hum.	Hum.	App.	Hum.	Hum.
τ	-0.09	-0.36	-0.36	-0.38	-0.36	-0.49
p-value	0.72	0.15	0.16	0.15	0.15	0.048

(b) Upper-body

Metric	Average jerk	Average accel.	Global CCA		Hellinger distance	FGD
Versus	Hum.	Hum.	Hum.	App.	Hum.	Hum.
τ	-0.11	-0.26	0.11	-0.49	-0.40	-0.51
p-value	0.64	0.27	0.64	0.041	0.085	0.029

Table 4.7 Rank correlations (Kendall's τ) between the "error" in the objective metrics (how much each objective value differed from the reference FNA/UNA) and median human-likeness scores (here abbreviated "Hum.") or – only for CCA – the preference for matched motion after splitting ties (abbreviated "App."). A strong predictor of human scores will exhibit a τ-value close to negative unity combined with a low p-value.

4.7 Discussion

4.7.1 Challenge Results

In this section we discuss the results from both editions of the GENEA Challenge, and look at parallels and differences between these editions.

When we look at the results for the first GENEA challenge, we can observe that no system was able to come close to the natural motion condition N. However, when comparing outcomes with previous baselines, it appears that already things have improved, and that progress is being made. For appropriateness, the difference, numerically, is larger between natural motion and machine-learning approaches, than for human-likeness. This could indicate that appropriateness of motion for the speech is a harder task to solve, and might simply require more than just a one-to-many mapping solution.

The results from the second GENEA challenge are different to the first one. The main differences between both challenges are that there are different teams participating, that there are two tiers they could submit to, and that the appropriateness of the motion for the speech is evaluated in a different way. When looking at the human-likeness scores, we can observe that each tier contains an entry that is rated significantly higher than the motion-capture recordings (sometimes also referred to as the ground-truth).

It is important to mention that the human-likeness evaluation is constrained by several factors. First, the recorded motion is not always an accurate depiction of real ground-truth motion. Especially when looking at finger motion, which was often not correctly recorded, and many teams excluded finger motion as part of their synthesis model. Secondly, the use of the avatar, lacking facial features, lowering the bandwidth, lowers the threshold for what can be seen as human-like.

In terms of appropriateness, no system beats the motion-capture recordings. It is clear that there is still a large gap in generating appropriate motion that matches to the speech.

4.7.2 Limitations

The challenges we run had several limitations, which we cover here. In the first challenge, the appropriateness methodology results implied a preference for systems that also scored high on human-likeness. We were able to successfully disentangle appropriateness and human-likeness in the second edition. This is visible in figures 4.10. We can see that there is a clear $x = y$ relationship for the results of GENEA 2020 on human-likeness and appropriateness, which was not the case two years later.

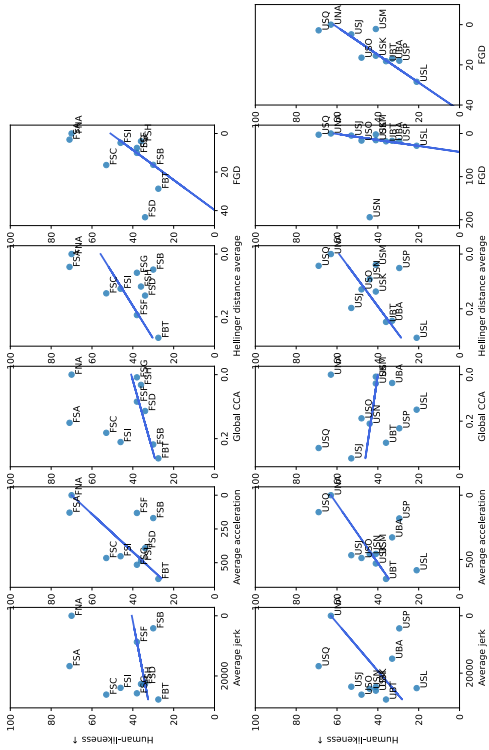


Figure 4.11 Scatterplots comparing objective metrics and human-likeness ratings. The first row is for the full-body tier and the second row is for the upper-body tier. The x-axis shows the absolute magnitude of the difference between the objective value for each system and the corresponding value for the reference motion FNA/UNA, with the scale reversed such that the systems most similar to the reference are on the right. Regression lines (from the Theil-Sen regressor (Sen, 1968; Theil, 1992), which is robust to outliers) are also shown. The last plot in the second row is for FGD but with a narrower x-axis range for a better view.

The appropriateness studies look at the appropriateness of the motion for the speech, however, in the first challenge semantic and rhythmic appropriateness were considered together. One way of addressing this would be to make the actual content of the speech inaudible, but to keep the rhythmic content of the speech. This could be one way of evaluating the appropriateness while zooming in on beat gestures.

For both challenges, the data we used was not perfect. Whereas the first challenge only utilised a single speaker, the second challenge had serious motion capture errors. Limbs could be unnaturally placed, or could be twitching. More diverse and higher quality data could improve challenge results. The audio stream was also missing certain information, as names and other contextual elements were often anonymised ('bleeped out'). Additionally, facial information was not part of the data set we used for the second challenge, which is a crucial element of nonverbal communication. Including facial and other nonverbal features in the evaluation would be a great improvement, at the cost of an increased complexity.

The gesture visualisation used in both challenges has several limitations. Some of these limitations are intentional, whereas others come with the data. For example, we came up with characters that lacked facial features. This was by design, to keep participants focused on the body poses. However, lacking a face significantly reduces the human-likeness of an avatar. We also rendered the 3D avatars to a 2D video with a fixed viewing point. Participants had no choice in how they could view the avatar. This gave us the freedom of clipping certain parts of the view, to obscure lower-body motion error, but limited the view for participants. One possible way of solving this would be instead provide 360 degrees videos in which participants could adjust the view of the avatar.

4.7.3 Evaluation of Challenges

Before we started the challenge series, we observed many approaches and evaluation techniques being used in the field of nonverbal behaviour generation. Hosting a challenge would mean we could have a meaningful impact on how results are evaluated and reported.

Whereas the first challenge relied on a single actor speaking freely about a variety of topics, for the second challenge, a dyadic dataset was used instead. Using a dyadic conversational dataset offers new possibilities but also poses new challenges. The second edition of the challenge was limited to motion generation of one person, while ignoring the rest of the conversation. This already brings in aspects of a dyadic conversation, but more implicitly. The step from a small, single speaker dataset, to a multi speaker dataset, was big. We believe that bringing in larger datasets, and shifting the task to include more speakers and interlocutors, would be beneficial in the long run. Although this increases computational complexity, it is necessary to focus on these aspects when we want ECAs to be able to interact with one or more persons. Another

aspect that should be brought forward is the inclusion of multiple nonverbal (facial) animations, since full body gesturing does not play a full role in the scope of nonverbal communication.

For the first challenge, a lot of infrastructure had to be written from scratch, or adapted from existing systems (such as HEMVIP (Jonell et al., 2021)). HEMVIP was modified for the second challenge, for example to automatically reject participants that failed a specific number of attention checks, or to include pairwise comparison studies. Some parts were revisited, such as the set up of the appropriateness studies or the avatar that was used. Although our idea was to set up a challenge once, and run it again given the framework that would be available from previous years, changing requirements and ideas often forced us to redo things. Both challenges were labour intensive for all involved (and a rough estimate would be an average of 2 to 3 months work per organiser, resulting in a 12 to 18 month time investment).

The way human-likeness was evaluated did not change over time, except for the inclusion of a training page at the second challenge. Although this way of measuring human-likeness is new (and not used by many others in the field), it provides promising results at a relatively low cognitive load. One idea for improvement could be to include variations of the human-likeness question, and treating human-likeness of motion as a construct that can come with multiple questions.

The evaluation of appropriateness of motion for the speech was changed significantly over time. Whereas the first challenge just posed a question on appropriateness, the second challenge relied on the (mis)matching paradigm. The reason for this was that the first challenge' results of appropriateness were tied to the human-likeness results. For the second challenge we decided to disentangle that, by using this new (mis)matching paradigm. Although assessing appropriateness on a scale from 0 to 100 provided a higher resolution, it also introduced a higher bias, something which is not the case with the new paradigm for appropriateness.

The current challenges only focus on human-likeness, in terms of motion quality, and appropriateness, of the motion for the speech. These two constructs are measured through on line user studies, with direct measuring methods. However, evaluating synthesised gesture motion should not be limited to human-likeness and appropriateness only. We should not only evaluate appropriateness of the motion of the speech, but also measuring other dimensions that need to be appropriate. One could think of appropriateness of generated motions for the affective state of a speaker, appropriateness for a given speaker, appropriateness for the context, and so on. Although evaluating all systems on an avatar in a human-avatar interaction such as done by he et al. (2022) is not feasible, both in time and complexity, we should definitely think of including other constructs that could be included and assessed directly (or indirectly depending on the instrument for measurement). For example, 'intelligibility' or 'understanding' often comes back as a construct when we look at the review results in chapter 3. Even when we find out that a certain

motion sequence is both human-like and appropriate for the speech, we do not know whether this motion sequence amplifies or weakens the message that is being conveyed. We therefor do not know whether it contributes to the intelligibility of the agent.

4.8 Conclusion

Over the past two years, we have hosted two editions of the GENEA Challenge. The second edition built upon and refined the first, improving both the data used and the evaluation methodologies applied. While the first challenge was part of the GENEA workshop, the second was included in the ICMI Grand Challenge series, significantly enhancing its impact.

Organizing these challenges has advanced the field's use and reporting of subjective methods. Future editions will build on the progress made so far. These challenges have established benchmarks for the current state-of-the-art, refined subjective evaluation methodologies, and provided valuable insights into the correlations between objective and subjective metrics.

Generating Nonverbal Behaviour

5

Generating Nonverbal Behaviour

5.1 Introduction

In this chapter we look at generating nonverbal behaviour for ECAs. The previous chapters mostly covered subjective evaluations for synthesised gestures, which triggered our interest to also look at the generation side.

We can easily say that gesture generation for ECAs has been picked up intensely by machine learning enthusiasts over recent years, resulting in an abundance of output that is concerned with model creation. On top of that, one could contribute with a new dataset, but this is a costly and difficult endeavour, that requires a fully equipped lab with motion capturing hardware and software. One of the dimensions that can still be explored and could yield novel findings, is the integration of multiple forms of nonverbal behaviour besides nonverbal speaking behaviour.

For this chapter, we did exactly that. We took an existing model, named StyleGestures (Alexanderson et al., 2020), adapted the model that it could handle new data structures, and compared its performance on the dataset used by the GENEA 2022 challenge, to a baseline model that was one of the better performing conditions of that challenge (see also chapter 4). We fed StyleGestures information from the dyadic conversation, such as the motion and audio signal from the interlocutor, to generate both speaking and listening behaviour. We used the models' output in five user studies where we looked at the human-likeness and appropriateness of both the speaking and listening motion, following the evaluation methodology from chapter 4.

5.2 Methods

In this section we describe the data and preprocessing, the models and the setup of the user studies for evaluating the generated speech and listening behaviour.

5.2.1 Data and Preprocessing

To ensure that the StyleGestures (SG) model is applicable to a wider range of conversational interactions, we opted to train it on a dataset that includes human dyadic interactions, rather than just a single speaker. Our data set of choice is "Talking With Hands 16.2," which provides a rich source of dyadic conversational data. This dataset includes both motion capture and audio, totalling 50 hours of recorded interactions. As the baseline model relies on text input for generating co-speech gestures, we made use of annotations provided by the GENEA Challenge 2022 (Yoon et al., 2022). More on this dataset can be found on page 71. For our experiments, we utilised a subset of 10 hours of conversation data from the combined data set. We opted for only including conversational takes that included the speaker labelled 'deep5' in the original data as a participant, since this was the single speaker with the most data in the data set. Furthermore, we conducted a thorough manual inspection of the data set to exclude takes that exhibited significant motion errors.

By adhering to these selection and inspection processes, we aimed to create a reliable and high-quality data set for training and evaluation purposes. The audio channel was transformed into a 27-channel mel-frequency representation following the original paper on SG (Alexanderson et al., 2020). The mel-frequency representation represents a spectrogram in which the distances in pitch sound equally distant to the listener, which is known as a mel-spectrogram. This in comparison to a normal spectrogram where this is not the case. The resulting features were down-sampled to 30 frames per second (FPS), to match up with the frame rate of the motion. Poses (joint rotations) were represented using exponential maps, which prevents discontinuities (Grassia, 1998), and full-body motion was used excluding finger and facial information. The input data to the model consisted of the concatenated audio features and speaker identity, and the motion of the interlocutor.

StyleGestures

The StyleGestures (SG) model (Alexanderson et al., 2020) is a probabilistic generative sequence model based on MoGlow, which uses normalizing flows (Henter et al., 2020; Papamakarios et al., 2021). The model was modified to accept dyadic input (speaker 1 and speaker 2), with the input being a concatenation of two audio streams, a one-hot encoding of the speaker identity, and the motion stream of the interlocutor (speaker 2). The output of the model was joint angles using the exponential map for speaker 1. The modified SG model was trained using the standard parameters from the SG paper, with a batch size of 120, noam_learning_rate_decay with 3000 warm up steps, and a minimum learning rate of 0.00015. The optimiser used was Adam, with a learning rate of 0.0015. Since the input data for this version of SG deals with dyadic information, the input size dimension is much larger than the original dimensions of the input data in Alexanderson et al. (2020), which only had

the audio features as input. Therefore, the model was trained for 160k steps before test motion was generated. We applied post-processing to the motion data to improve the quality of our generated listening behaviour. Specifically, we used a Butterworth lowpass filter to smooth the rotation data and filter out minor motion glitches. The cutoff frequency was set to 3.0 Hz and the filter order was set to 4, as we found these values to work the best with our output data. We conducted user studies to compare the output of this model to the recorded ground-truth motion.

Baseline

We wanted to compare our results to a model that had already been applied to the data set we used. For this, we selected the "The IVI Lab entry to the GENEA Challenge 2022", since the code for this entry was openly available and tested by others, winning the reproducibility award at the challenge (Chang et al., 2022). The baseline model is based on the Tacotron2 architecture that is used for speech synthesis with a locality constraint attention mechanism, and takes text and speech audio as input to generate motion data (Shen et al., 2018). It was trained on only the text and speech input data from the speaker whose motion we are predicting, namely speaker 1 (in contrast to our SG model that was trained on full dyadic data). For the training parameters we relied on the values used by Chang et al. (2022).

5.2.2 Visualisation

We rendered the generated motion on a faceless avatar, which we used before in the GENEA 2022 challenge, and can be seen in Figure 4.2b.

5.2.3 User Studies

We designed five user studies to evaluate the performance of our adapted model. For the first study, we relied on the appropriateness methodology to see whether participants could pick out motion sequences in which the avatar appeared to be listening. For study two we looked at human-likeness for gesticulation. Study three investigated the appropriateness dimension for gesticulation. In study four, we looked at human-likeness for listening behaviour, and in study five we looked at the appropriateness of the generated listening behaviour.

Study 1: ``Does it listen?''

The purpose of this study was to investigate the ability of participants to identify generated listening segments when presented with unrelated speech

motion fragments. We recruited 32 participants who were required to be native English speakers. The listening segments were generated using either the baseline model or the SG model. To determine whether participants were able to distinguish matching listening motion versus mismatching speech motion, speech motion segments were obtained from the ground-truth. Each matching or mismatching segment was then added to a video containing a speaker, who was positioned on the left of the video with the listener on the right. The audio for each conversation was added to the video. In total, we selected 30 listening segments per condition, totalling 60 segments. The videos containing the conversations (matching versus mismatching) were presented side by side in a random order, and the order of presentation was also randomised. Participants were asked the question: "Please indicate in which of the two clips the character on the right moves like a listening person." The interface for the user study followed the one that was designed by Wolfert et al. (2021) for their study that involved pairwise comparisons. Participants had the choice between three options: the left video, the right video, or both are equal. Each participant was presented two attention checks, inserted at random points during the experiment. One attention check was text based and the other one audio based, halfway the video it would ask the the participant to select the button belonging to that specific video. We used Barnard's test for identifying statistically significant differences between conditions at the level of $\alpha = 0.05$, as it is more powerful than Fisher's exact test when testing two independent binomials. Next to that, the Holm-Bonferroni method was applied to correct for multiple comparisons.

Study 2: ``Human-likeness for gesticulation''

For this study we used the same approach for the interface and the attention checks. The purpose for this study was to investigate how human-like the speaking behaviour was, that was generated by the model. For this, we compared it to the baseline and the ground-truth motion. We recruited 22 participants who were required to be native English speakers. From the test set, 30 segments were selected in which the avatar was talking, and gesture motion was synthesised from SG and the baseline, or taken from the ground-truth. Participants were asked the following question: *"How human-like does the gesture motion appear?"* They were asked to rate the human-likeness on a scale from 1 to 100, where a score of 100 would mean the gesture motion was completely human-like. Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

Study 3: Appropriateness for Gesticulation

For this study, we examined the appropriateness of gesture motion for the speech generated by the model. We followed the appropriateness paradigm introduced by Rebol et al. (2021) in which matching and mismatching stimuli

are put on one screen side-by-side. We recruited 27 participants. To form our stimuli, we took the same 30 segments used in study 1 and chose 30 additional segments as mismatching stimuli. These segments were then paired with the interlocutor, resulting in two avatars being visible in each video. The speaker was placed on the left side, whereas the interlocutor was placed on the right. For each of the 30 videos, we provided a mismatching video with motion unrelated to that part of the conversation. These videos were paired with the matching interlocutor. To establish an appropriateness baseline, we included matched and mismatched videos from the ground truth. We hypothesised that participants would be able to identify the correct segments for direct motion-captured gesticulation. Both videos were placed on the same page, and participants were asked to indicate in which of the two clips the character on the left moves appropriately for the speech. The interface for the user study followed the one that was designed by Wolfert et al. (2021) for their study that involved pairwise comparisons. Participants had the choice between three options: the left video, the right video, or both are equal. Throughout the experiment, each participant encountered two attention checks, inserted at random places during the experiment. One attention check was text-based and the other one audio based, halfway through the video it would ask the participant to select the button belonging to that specific video. We used Barnard's test for identifying statistically significant differences between conditions at the level of $\alpha = 0.05$. Additionally, we applied Holm-Bonferroni to correct for multiple comparisons.

Study 4: ``Human-likeness for listening''

The purpose for this study was to investigate how human-like the generated listening behaviour was. For this, we compared it to the baseline and ground-truth motion. We recruited 22 participants who were required to be native English speakers. From the test set, 30 listening segments were selected, and listening motion was synthesised from SG and the baseline, or taken from the ground-truth. The videos did not feature audio, as we wanted participants to specifically focus on the motion. Participants were asked the following question: *"How human-like does the listening motion appear?"*, and had to rate the videos on a scale from 1 to 100. Three videos were placed on one screen, using the HEMVIP framework for evaluating the stimuli (Jonell et al., 2021). The order of the videos on the screen was randomised, as well as the order in which the screens were presented to the participant. Each participant was presented with two attention checks, inserted at random places during the experiment. Both attention checks would ask the participant to rate the video with a certain score. The text for the attention check would only appear halfway the video. Each participant rated 14 screens with 3 stimuli per screen, totalling 42 ratings per participant and 308 ratings per condition.

Study 5: Appropriateness for Listening

For this study, we aimed to investigate the appropriateness of the listening motion for the conversation, generated by the model. We recruited 27 participants. The setup of this study followed the setup for study 2, but instead of selecting speaking segments, we selected segments where the main speaker was listening to the interlocutor. We took 30 segments, and 30 additional segments as mismatching stimuli. These segments were then paired with the other speaker, resulting in two avatars being visible side-by-side in each video. Now, the listener, for which the motion was synthesised, was placed on the left. Both videos were placed on the same page, and participants were asked to indicate in which of the two clips the character on the left moved appropriately for the speech. To establish an appropriateness baseline, we included matched and mismatched videos from the ground truth. We hypothesised that participants would be able to identify the correct segments for full-body listening behaviour. Throughout the experiment, each participant was presented two attention checks, inserted at random places during the experiment. One attention check was text based and the other one audio based, halfway through the video a text would appear or an audio message could be heard asking the participant to select a specific option in the interface. We used Barnard's test for identifying statistically significant differences between conditions at the level of $\alpha = 0.05$. Additionally, we applied Holm-Bonferroni to correct for multiple comparisons.

5.2.4 Objective Analysis

As pointed out often before, there is no single objective metric that can capture the quality of the generated motion. Therefore, we rely on commonly used objective metrics in the field such as the acceleration, jerk and velocity histograms.

5.3 Results

5.3.1 User Studies

In this section we describe the results of the user studies we did for the evaluation of the two models on speech and listening behaviour generation.

Study 1: ``Does it listen?''

In this study, we looked at matching versus mismatching for listening behaviour, where the mismatched video used unrelated speech motion. Participants were presented with pairs of matching/mismatching videos and asked

Study	N	Mean Age (SD)	Male	Female	Nationality
1	30	40.6 (11.5)	15	15	UK (28), USA (1), NZ (1)
2	22	35.2 (12.4)	16	6	UK (20), USA (1), IE (1)
3	27	40 (12.54)	14	13	UK (21), CA (2), IE (3), AU (1)
4	22	41.8 (13.94)	13	9	UK (16), USA (1), IE (2), NZ (2), CAN (1)
5	26	39 (11.55)	10	16	UK (15), CA (5), IE (3), AU (3)

Table 5.1 Participant demographics for each study.

to choose which one featured the listening motion. They also had the option to choose that the videos were equal.

For SG 46 (16%) videos were reported as equal, 178 (61%) as matching and 69 (23%) as mismatched. For the baseline this was 25 (8%) reported as equal, 215 (72%) as matching and 57 (20%) as mismatching. We further performed Barnard's test with Holm-Bonferroni correction to analyse the data. In the SG condition, we found a significant difference between matched and mismatched videos (Chi2 stat: 69.0, p-value: < 0.001). Similarly, in the baseline condition, there was a significant difference between matched and mismatched videos (Chi2 stat: 57.0, p-value < 0.001). These results suggest that participants were able to perceive which video of a pair featured the listening behaviour.

Study 2: ``Human-likeness for gesticulation''

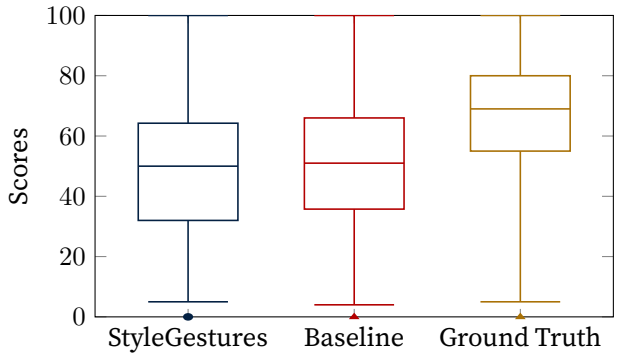


Figure 5.1 Boxplots of human-likeness scores on gesturing for StyleGestures (SG), baseline (BL), and ground truth conditions (GT).

For this study we explored the construct of human-likeness for gesticulation.

The median score for SG was 47 (95% CI[45.00,49.00]), for the baseline 41(95% CI[40.00, 44.00] and for the ground truth 56.5(95% CI[53.00, 60.00]. We conducted Wilcoxon signed-rank tests to the SG, baseline and ground-truth conditions. It revealed that there was a significant difference in the similarity ratings between the SG and ground-truth conditions (W=6116.0, p<0.001) as well as between the baseline and ground-truth conditions (W=6865.5, p<0.001). However, there was no significant difference in the similarity ratings between the SG and baseline conditions (W=20631.0, p=0.097).

Study 3: ``Appropriateness for Gesticulation''

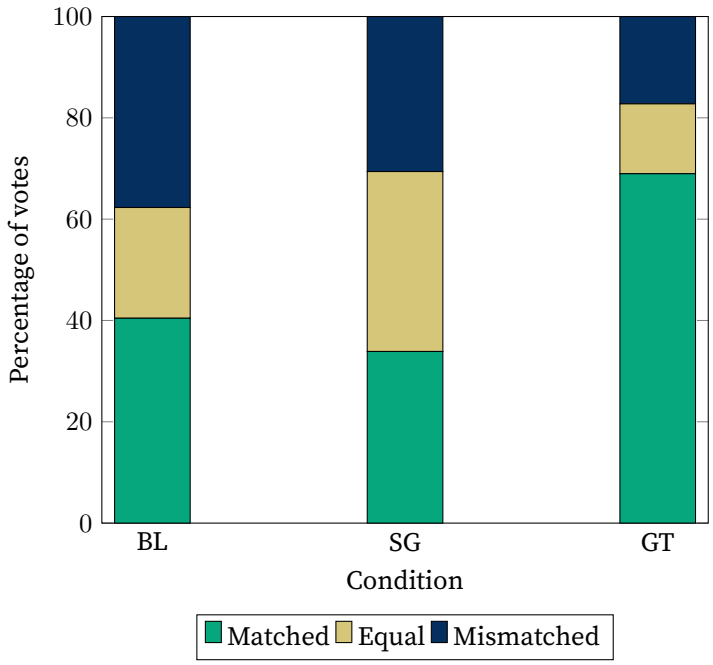


Figure 5.2 Stacked bar charts showing the percentage of votes on gesturing for StyleGestures (SG), baseline (BL), and ground truth conditions (GT) in study 2.

The purpose of this study was to investigate the ability of participants to select the correct, matching, segment belonging to a conversation. For this, we used the match/mismatch paradigm initially proposed by Rebol et al. (2021), later also used by Yoon et al. (2022). Participants were presented with pairs of matching/mismatching videos and asked to choose which one featured the correct gesturing motion. They also had the option to choose whether the videos were equal. Details on demographics can be found in Table 5.1.

Figure 5.2 shows the percentage of votes for matched, equal and mismatched per condition. For SG 62 videos were reported as matching, 56 as mismatched and 65 as equal. For the baseline condition, 74 were reported as matching, 69 as mismatching and 40 as equal. For the ground truth, 120 were matched, 30 were mismatched and 24 were reported as equal.

To analyse these results, chi-square tests were conducted with Holm-Bonferroni correction applied for multiple comparisons. For SG, matching differed significantly from mismatching ($\chi^2 = 179, p < 0.0001$). For baseline, matching differed significantly from mismatching $\chi^2 = 179, p < 0.0001$, as well as for ground truth ($\chi^2 = 155, p < 0.0001$).

Lastly, we tested for differences between the conditions, ties were split equally over matching and mismatching. For SG versus ground-truth, there was a significant difference ($\chi^2 = 21.99, p < 0.0001$. For baseline versus ground-truth, there was a significant difference *$\chi^2 = 21.99, p < 0.0001$. SG and baseline did not differ significantly.

Study 4: ``Human-likeness for listening''

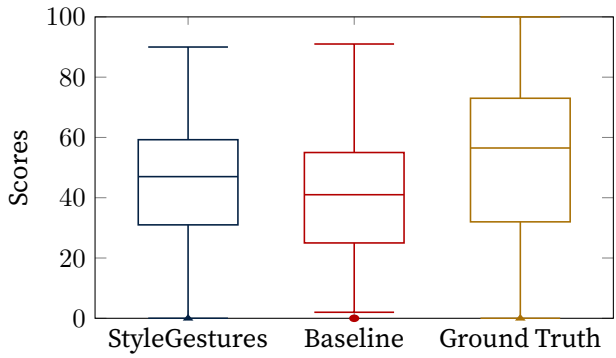


Figure 5.3 Boxplots of human-likeness scores for listening behaviour.

This study examined the human-likeness for listening behaviour of SG compared to a baseline and the ground-truth (GT). 22 participants were recruited, all participants passed the attention checks. Of these, the mean age was 41.8 years (SD=13.94). 9 identified as female and the other 13 identified as male. 16 participants were from the UK, 1 from the USA, 2 from Ireland, and 2 from New Zealand.

The median score for SG was 47 (95% CI[45.00,49.00]), for the baseline 41(95% CI[40.00, 44.00] and for the ground truth 56.5(95% CI[53.00, 60.00]. To further analyse the data, Wilcoxon signed-rank tests were conducted between SG and baseline, SG and GT, and baseline and GT conditions for listening. The results showed that there was a statistically significant difference in the human-likeness perception between the SG and baseline conditions (Z = 16265.5, p

< 0.0001). The results also showed a significant difference between the SG and GT conditions (Z = 16506.0, p < 0.0001). Lastly, there was a significant difference between the baseline and GT conditions (Z = 11646.5, p < 0.0001).

Study 5: ``Appropriateness for Listening''

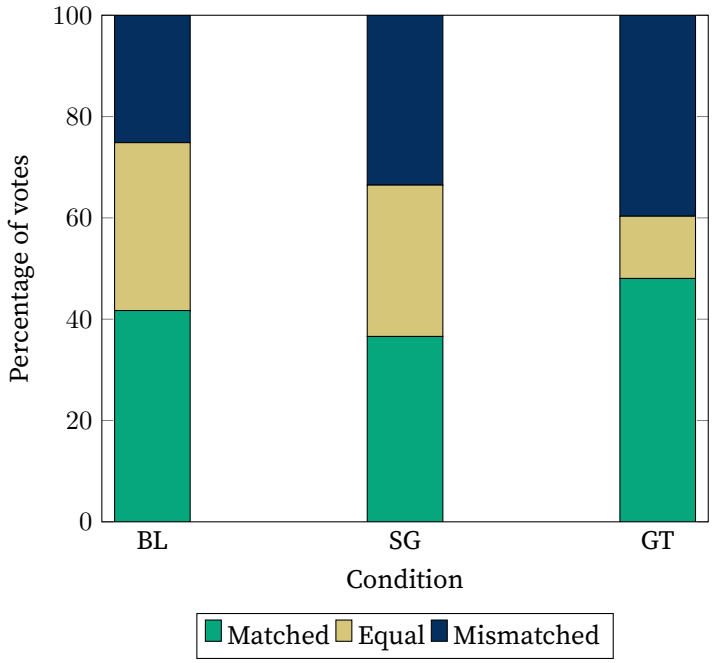


Figure 5.4 Stacked bar charts showing the percentage of votes on listening for Baseline (BL), StyleGestures (SG) and Ground Truth (GT) in study 4.

Figure 5.4 shows the percentage of votes for matched, equal and mismatched per condition. For SG 60 videos were reported as matching, 66 as mismatched and 49 as equal. For the baseline condition, 73 were reported as matching, 44 as mismatching and 58 as equal. For the ground truth, 86 were matched, 71 mismatched and 22 reported as equal.

For SG, matching differed significantly from mismatching ($\chi^2 = 149, p < 0.0001$). For baseline, matching differed significantly from mismatching $\chi^2 = 170, p < 0.0001$, as well as for ground truth ($\chi^2 = 174, p < 0.0001$).

Lastly, we tested for differences between the conditions, ties were split equally over matching and mismatching. There were no significant differences between the three conditions. Details on demographics can be found in Table 5.1.

Condition	Mean Jerk	Mean Acceleration
Ground-truth S	38660.78 (SD=830	1101.37 (SD=287.00)
Ground-truth L	23980.68 (SD=4494.39)	524.74 (SD=148.00)
Baseline S	10318.88 (SD=2741.81)	422.34 (SD=120.20)
Baseline L	4633.45 (SD=1981.94)	182.61 (SD=84.49)
StyleGestures S	3392.73 (SD=6620.53)	235.97 (SD=306.05)
StyleGestures L	3395.10 (SD=4340.59)	215.46 (SD=174.55)

Table 5.2 Mean Jerk and mean Acceleration for the generated speech (S) and listening (L) behaviour.

5.3.2 Objective Analysis

We calculated the mean jerk and mean acceleration as well as the velocity histograms for the three conditions (and specified for listening and speech). Velocity histograms depict the distribution of gesture speeds in generated speech motion. They offer insights into the naturalness and fluidity of gestures, The result for the listening and speech motion can be found in table 5.2. The velocity histograms are visualised in figure 5.5 for listening, and figure 5.6 for speech.

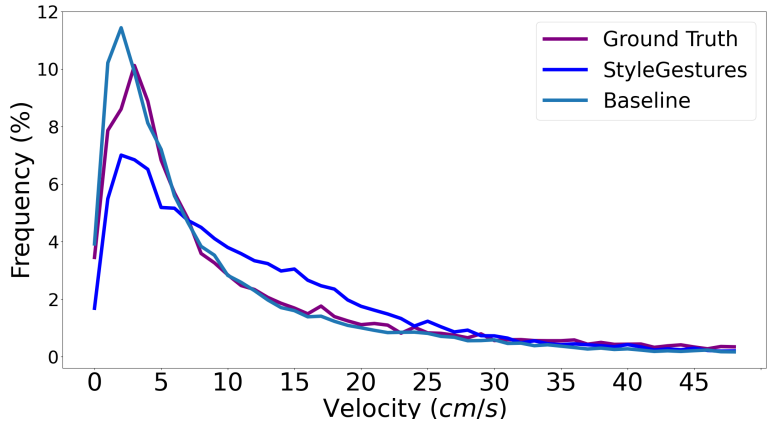


Figure 5.5 Velocity histogram for the listening behaviour test samples.

5.4 Discussion

We conducted five user studies to evaluate the quality of our model on generating listening motion and speech gestures. We found that our model under

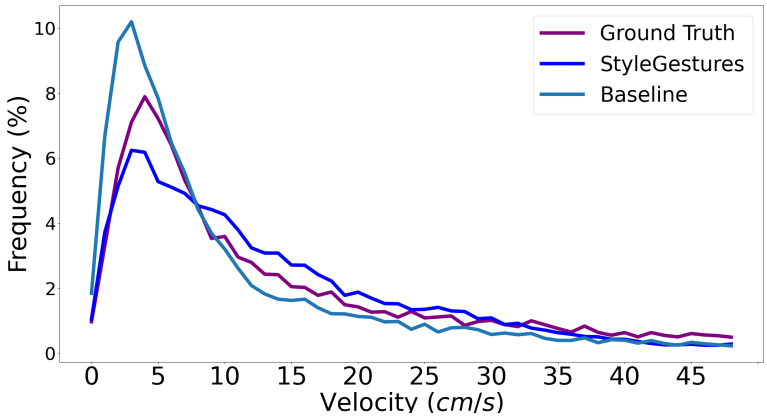


Figure 5.6 Velocity histogram for the speech behaviour test samples.

performs in comparison to the baseline and the ground-truth for the mismatching study. Even though 60% are correctly identified as matching stimuli, more stimuli are identified as "they're equal", than for the baseline condition. It shows that for quite some situations participants found it hard to identify the correct segment. This was also the case for the baseline model where only 72% was correctly identified as matching stimuli.

In the second study we evaluated human-likeness for speaking. We found a significant difference for the two conditions with the ground-truth, but no significant difference between SG and baseline, which is an interesting finding since the baseline model has the advantage of incorporating semantic information in relation to its gestures. However, the notion of semantic related gestures is not something that we can catch with human-likeness evaluations, since these revolve around motion quality and not appropriateness of gestures with speech audio.

For the third study we evaluated the appropriateness of the generated motion for speaking. Which we did through the use of matching and mismatching videos (Rebol et al., 2021). In study 3 we took the same 30 segments and combined them in one video with the interlocutor. In one of the two videos presented to the participant, the gesture sequence of the avatar on the left was not related to that part of the conversation. We found significant differences for both baseline and SG with the ground-truth condition, but not for baseline versus SG. When we look at 5.2 we see that more videos are identified as 'equal' for the SG condition. As expected, ground truth videos are identified as matching for more than 70% of the time. Since the baseline model has also access to text besides speech audio, one would expect this model to generate more appropriate (and even semantic related) gestures, but the results from the appropriateness study do not seem to confirm that.

We wanted to know whether these two paradigms of human-likeness and

appropriateness testing, could be used with more subtle forms of human nonverbal behaviour, such as listening behaviour. In the fourth study we evaluated the human-likeness of the generated listening behaviour and compared it to the ground-truth. It is important to mention that for the human-likeness evaluation we excluded the audio, to only assess the quality of the motion. We found significant differences between all conditions, with SG scoring after the ground truth. However, the overall rating for each condition was not very high. We think that human-likeness testing for motion for listening behaviour is difficult since appropriate listening behaviour is really dependent on the conversation. Omitting the audio could also have led to the participants not being able to see that this motion is supposed to be part of a conversation. Another reason could be that in terms of motion quality it all was similar, and therefore scored the same because of the lack of context.

For the last study we evaluated the appropriateness for the generated listening behaviour. Appropriateness testing of listening behaviour could help figure out whether it actually matters what listening behaviour is tied to a conversation, and whether participants can spot differences in generated listening behaviour. Here, we cannot report any significant difference between the three conditions. It appears that participants have difficulty identifying the right listening behaviour. One reason for that could be that listening behaviour takes place more with facial expressions than with body language, and that body pose alone is not enough to say that someone is attending a conversation. Another reason could be that the avatar visualisation is too far from human-likeness, and therefor participants have a harder time believing that it is a human that is moving. Listening behaviour is not only dependent on full-body motion, but is also often combined with verbal feedback (Gómez Jáuregui et al., 2021).

When we look at the results of the objective metrics, we can observe that for the baseline the jerk and acceleration is much higher for the speech behaviour than the listening behaviour. For SG, there is not a large difference in mean jerk and mean acceleration between speech and listening (although the standard deviation is). We see this pattern confirmed in the velocity histograms, here the ground-truth and SG are closer to each other.

Since the main aim of this work was to compare SG with the baseline model and the ground-truth for generated listening behaviour, the results from study 1 and 2 give an indication that we can use generative models, originally used for co-speech gesture generation, for generating listening motion.

5.5 Conclusion

In this chapter we looked at the effectiveness of applying a generative model for motion generation to generating listening motion. We compared three conditions against one and other, and found that our approach comes close to the ground-truth for listening motion for human-likeness. Our approach is

a first good step in the direction of integrating full-body nonverbal feedback in an automated and generative way using one model only. However, further research is necessary to see how well the results of our model work in real life interactive scenarios.

Conclusion & Future Perspectives

6

Conclusion & Future Perspectives

This chapter provides a summary of the research conclusions in section 6.1, and future perspectives in 6.2.

6.1 Summary

This thesis investigated the evaluation and data-driven generation of co-speech gestures in ECAs. We first discussed the necessary background and reviewed the subjective evaluation methods used in the field. Following our results, we decided to look into a variety of evaluation methods, and use them ourselves. In one situation we compared to hand made gestures, and in another situation we used existing computer generated stimuli. We looked at whether the field should keep on using 5-point scales, or whether they should solely choose for pairwise comparisons. Finally, we decided to also test out a questionnaire, something we suggested to use at the end of our review in Chapter 2. We also co-organised two gesture generation challenges, where we crowdsourced user evaluations. This was done for two consecutive years, which provided a lot of new insights. Finally, we turned our gaze towards producing speech and listening behaviour, for which we adapted an existing model to work with dyadic conversational data, and which we also compared to a baseline model from one of the two challenges.

6.1.1 On Subjective Evaluation Methods

Our review in Chapter 2 found that many studies failed to report details on participant demographics. Additionally, it was found that many studies do not report on the inclusion of ground truth motion, and that there is often not a open access baseline included in their comparison studies. Finally, many different questionnaire items, albeit with some overlap, have been used in recent years, to evaluate stimuli. Following these findings, we suggested a list of items that are preferred to be reported, as well as a set of sentences and scenarios that could be used in the future for evaluating generated co-speech gestures.

6.1.2 Comparing and Evaluating Gestures

Chapter 3 reported on three studies that covered four different evaluation approaches. We found that using a ranking approach did not provide that much information, and if more information is needed than what a ranking approach could provide, it is better to use rating scales. However, when rating scales were compared to pairwise comparisons, both had their pros and cons. Finally, we tested out a questionnaire, but even though we expected this to provide significant results, we had to conclude that on each of the three constructs, there were no significant differences between the included conditions.

6.1.3 Crowdsourcing Gesture Generation Systems and Evaluations

Chapter 4 reported on two editions of the GENEA Challenge. Each challenge had its specifics, but the task and the aim were the same: given a dataset, generate nonverbal behaviour, and we will do the evaluation. When we look at the evaluation part, we can conclude that the disentanglement between appropriateness and human-likeness is successful for the second edition of the challenge, in contrast to the first edition, where the scores for human-likeness and appropriateness were much more alike. For the second edition, there was even a participating system that scored higher on human-likeness than the original recorded motion.

6.1.4 Generating Nonverbal Behaviour

The last chapter was concerned with generating nonverbal behaviour, and more specifically, co-speech gestures and listening behaviour. For this we made use of an existing dataset and model, but adapted the model significantly to work with dyadic conversational data. As we were also interested in generating listening behaviour, we applied several user studies. The information provided by the user studies differed, more so due to the subtle nature of the included listening behaviour. When using the existing (or sometimes slightly adapted) paradigms for evaluating human-likeness and appropriateness, we come across the problem of usability for the more subtle forms of nonverbal behaviour such as listening behaviour. Altogether we found that it is feasible to incorporate not only speech behaviour in the training set, but also listening behaviour, and that it is possible to generate both using only one model. Further research is needed to verify these results in an interactive setting.

6.2 Future Perspectives

Much of the research described in this thesis is concerned with evaluation methods, and how these could be improved. Besides evaluation methodology, there are also other aspects of co-speech gesture generation that need specific attention for future improvements.

6.2.1 Datasets

There are still not many available datasets, the overview in Nyatsanga et al., 2023 lists 17 English speaking datasets (19 if we include the adapted datasets by GENEA). The majority of the listed datasets are monologues, and do not include finger data. Besides that, not all datasets contain 3D motion data, which is the standard for data needed to train new models for generating co-speech gestures.

Another issue is that most of this data is not annotated. The communicative intent is often not included, hence the reason that much work on data-driven gesture generation up till now is based on a one to many mapping.

For the progress in the field to really take off, it is necessary that new datasets arise. One thing is that there needs to be more (cultural and linguistic) diversity. Applying gesture generation models on ECAs that are not equipped for specific regions could result in showing nonverbal behaviour that is inappropriate for the context an ECA is in.

6.2.2 Models

There has been a lot of development in terms of models for co-speech gesture generation, and many of the latest models not only have audio and text as input, but also other features such as speaker id and affective state. Including these features, and also being able to include the conversational partner(s), makes that these models are modelled to do what we expect them to do: mapping verbal input to nonverbal behaviour. However, in contrast to work from the early 2000s, most of these models do not take into account the communicative intent. And this is an aspect that needs improvement, but is at the same time also a hard task.

6.2.3 Evaluation Paradigms

In this work we have covered a variety of evaluation paradigms, and each one of them has their pros and cons. The current approach for appropriateness testing seems to work better for assessing the appropriateness of the motion for the speech. However, these evaluations can be improved. he et al., 2022 evaluated an existing model in a virtual avatar in an interaction. Here, only

behavioural measurements such as gaze yielded significant differences. Still, evaluating in an interactive scenario should be preferred, as this is the aim for this field: equipping ECAs with the capacities to communicate with humans in a natural and convincing way. Another improvement could be made for the current one-way method of evaluating stimuli: ablation studies to look at the effect of body pose, gestures and facial expressions.

6.3 Final Remarks

Central to this thesis was the question: ”Can we improve and advance the standard of subjective evaluations for the field of nonverbal behaviour generation?”. We did so through reviewing subjective evaluation methods, comparative studies, the organisation of two challenges and finally the generation of nonverbal (listening) behaviour ourselves. Through these diverse approaches, our aim was not solely to enhance the quality of subjective evaluations but also to push the boundaries of knowledge and practice within the field of nonverbal behavior generation, contributing to its ongoing advancements.

References

Alexanderson, S. (2020). The stylegestures entry to the genea challenge 2020. *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088600

Alexanderson, S., Henter, G. E., Kucherenko, T., & Beskow, J. (2020). Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum, 39*(2), 487–496. https://doi.org/10.1111/cgf.13946

Alexanderson, S., Nagy, R., Beskow, J., & Henter, G. E. (2023). Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph., 42*(4), 1–20. https://doi.org/10.1145/3592458

Allmendinger, K. (2010). Social presence in synchronous virtual learning situations: The role of nonverbal signals displayed by avatars. *Educational Psychology Review, 22*(1), 41–56. https://doi.org/10.1007/s10648-010-9117-8

Aly, A., & Tapus, A. (2013). A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 325–332. https://doi.org/10.1109/HRI.2013.6483606

Amioka, S., Janssens, R., Wolfert, P., Ren, Q., Pinto Bernal, M. J., & Belpaeme, T. (2023). Limitations of audiovisual speech on robots for second language pronunciation learning. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 359–367. https://doi.org/10.1145/3568162.3578633

Ao, T., Zhang, Z., & Liu, L. (2023). Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Trans. Graph., 42*(4). https://doi.org/10.1145/3592097

Barnard, G. A. (1945). A new test for 2 × 2 tables. *Nature, 156*, 783–784. https://doi.org/10.1038/156783b0

Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., & Šabanović, S. (2020). *Human-robot interaction: An introduction*. Cambridge University Press.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*, 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bennewitz, M., Faber, F., Joho, D., & Behnke, S. (2007). Fritz - a humanoid communication robot. *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 1072–1077. https://doi.org/10.1109/ROMAN.2007.4415240

Bergmann, K., & Kopp, S. (2009). Gnetic – using bayesian decision networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Intelligent virtual agents. iva 2009. lecture notes in computer science* (pp. 76–89, Vol. 5773). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04380-2_12

Bergmann, K., Kopp, S., & Eyssel, F. (2010). Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents. iva 2010. lecture notes in computer science* (pp. 104–117, Vol. 6356). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15892-6_11

Boersma, P., & Weenink, D. (2002). Praat v. 4.0. 8. *A system for doing phonetics by computer. Institute of Phonetic Sciences of the University of Amsterdam*, *218*, 1–2.

Bosker, H. R., & Peeters, D. (2021). Beat gestures influence which speech sounds you hear [Original work published January 27, 2021]. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943). https://doi.org/10.1098/rspb.2020.2419

Boynton, R. M. (1989). Eleven colors that are almost never confused. *Human Vision, Visual Processing, and Digital Display*, *1077*. https://doi.org/10.1117/12.952730

Bozkurt, E., Erzin, E., & Yemez, Y. (2015). Affect-expressive hand gestures synthesis and animation. *2015 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. https://doi.org/10.1109/ICME.2015.7177478

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. http://www.jstor.org/stable/2334029

Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 708–713. https://doi.org/10.1109/IROS.2005.1545011

Bremner, P., Pipe, A. G., Fraser, M., Subramanian, S., & Melhuish, C. (2009). Beat gesture generation rules for human-robot interaction. *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, 1029–1034. https://doi.org/10.1109/ROMAN.2009.5326136

Bremner, P., Pipe, A., Melhuish, C., Fraser, M., & Subramanian, S. (2009). Conversational gestures in human-robot interaction. *2009 IEEE International Conference on Systems, Man and Cybernetics*, 1645–1649. https://doi.org/10.1109/ICSMC.2009.5346903

Burton, N., Burton, M., Rigby, D., & et al. (2019). Best-worst scaling improves measurement of first impressions. *Cognitive Research, 4*, 36. https://doi.org/10.1186/s41235-019-0183-2

Buschmeier, H., & Kopp, S. (2018). Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1213–1221.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., Yan, H., et al. (2000). Human conversation as a system framework: Designing embodied conversational agents. *Embodied conversational agents*, 29–63.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 413–420.

Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. *Proc. SIGGRAPH*, 477–486.

Chang, C.-J., Zhang, S., & Kapadia, M. (2022). The ivi lab entry to the genea challenge 2022 – a tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 784–789. https://doi.org/10.1145/3536221.3558060

Chidambaram, V., Chiang, Y.-H., & Mutlu, B. (2012). Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 293–300. https://doi.org/10.1145/2157689.2157798

Chiu, C.-C., & Marsella, S. (2011). How to train your avatar: A data driven approach to gesture generation. In H. H. Vilhjálmsson, S. Kopp, S. Marsella, & K. R. Thórisson (Eds.), *Intelligent virtual agents. iva 2011. lecture notes in computer science* (pp. 138–151, Vol. 6895). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23974-8_14

Chiu, C.-C., & Marsella, S. (2014). Gesture generation with low-dimensional embeddings. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 781–788.

Chui, K. (2005). Topicality and gesture in chinese conversational discourse. *LANGUAGE AND LINGUISTICS-TAIPEI-*, *6*(4), 635.

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? using the "elochoice" package to assess pairwise comparisons of perceived physical strength. *PLOS ONE*, *13*(1), e0190393. https://doi.org/10.1371/journal.pone.0190393

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1), 61–104. https://doi.org/10.1177/147078530805000106

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods, 14*(4), 349–366. https://doi.org/10.1037/a0016956

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science, 4*, 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Elliott, L. L. (1958). Reliability of judgments of figural complexity. *Journal of Experimental Psychology, 56*(4), 335–338. https://doi.org/10.1037/h0043971

Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., & Alías, F. (2014). Gesture synthesis adapted to speech emphasis. *Speech Communication, 57*, 331–350. https://doi.org/https://doi.org/10.1016/j.specom.2013.06.005

Ferstl, Y., & McDonnell, R. (2018). Investigating the use of recurrent motion modelling for speech gesture generation. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 93–98. https://doi.org/10.1145/3267851.3267898

Ferstl, Y., Neff, M., & McDonnell, R. (2019). Multi-objective adversarial gesture generation. In *Proceedings of the 12th acm siggraph conference on motion, interaction and games*. Association for Computing Machinery. https://doi.org/10.1145/3359566.3360053

Fitrianie, S., Bruijnes, M., Richards, D., Abdulrahman, A., & Brinkman, W.-P. (2019). What are we measuring anyway? - a literature survey of questionnaires used in studies reported in the intelligent virtual agent conferences. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 159–161. https://doi.org/10.1145/3308532.3329421

Fitrianie, S., Bruijnes, M., Richards, D., Bönsch, A., & Brinkman, W.-P. (2020). The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3383652.3423873

Gałecki, A., & Burzykowski, T. (2013). Linear mixed-effects model. In *Linear mixed-effects models using r* (pp. 245–273). Springer.

Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2019). Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. *Advanced Robotics, 33*(7–8), 325–337. https://doi.org/10.1080/01691864.2019.1589570

Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2018). The influence of social cues in persuasive social robots on psychological reactance and compliance. *Computers in Human Behavior, 87*, 58–65. https://doi.org/10.1016/j.chb.2018.05.016

Gillies, M., Pan, X., Slater, M., & Shawe-Taylor, J. (2008). Responsive listening behavior. *Computer Animation and Virtual Worlds*, *19*, 579–589. https://doi.org/10.1002/cav.267

Gómez Jáuregui, A., Giraud, T., Isableu, B., & Martin, J.-C. (2021). Design and evaluation of postural interactions between users and a listening virtual agent during a simulated job interview. *Computer Animation and Virtual Worlds*, *32*, e2029. https://doi.org/10.1002/cav.2029

Grassia, F. S. (1998). Practical parameterization of rotations using the exponential map. *Journal of Graphics Tools*, *3*(3), 29–48. https://doi.org/10.1080/10867651.1998.10487493

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (Fourth). Advanced Analytics.

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners* (Vol. 92). John Wiley & Sons.

Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal communication. *Annual Review of Psychology*, *70*(Volume 70, 2019), 271–294. https://doi.org/https://doi.org/10.1146/annurev-psych-010418-103145

Ham, J., Cuijpers, R. H., & Cabibihan, J.-J. (2015). Combining robotic persuasive strategies: The persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics*, *7*, 479–487. https://doi.org/10.1007/s12369-015-0280-4

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., & Sumi, K. (2018). Evaluation of speech-to-gesture generation using bi-directional lstm network. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 79–86. https://doi.org/10.1145/3267851.3267878

he, Y., Pereira, A., & Kucherenko, T. (2022). Evaluating data-driven co-speech gestures of embodied conversational agents through real-time interaction. *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3514197.3549697

Henter, G. E., Alexanderson, S., & Beskow, J. (2020). MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph.*, *39*(4), 236:1–236:14. https://doi.org/10.1145/3414685.3417836

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., & Schröder, M. (2011). Generating listening behaviour (R. Cowie, C. Pelachaud, & P. Petta, Eds.). https://doi.org/10.1007/978-3-642-15184-2_17

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, *26*(1), 4–27. https://doi.org/10.1177/0261927X06296428

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. http://www.jstor.org/stable/4615733

Hömke, P., Holler, J., & Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0208030

Huang, C.-M., & Mutlu, B. (2013). Modeling and evaluating narrative gestures for humanlike robots. *Robotics: Science and Systems IX*. https://doi.org/10.15607/rss.2013.ix.026

Huang, C.-M., & Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 57–64. https://doi.org/10.1145/2559636.2559668

Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3-to 5-year-old children. *Journal of Experimental Child Psychology, 156*, 99–112.

International Telecommunication Union, Telecommunication Standardisation Sector. (1996). *Methods for subjective determination of transmission quality* (Recommendation No. ITU-T P.800). https://www.itu.int/rec/T-REC-P.800-199608-I

Ishi, C. T., Machiyashiki, D., Mikata, R., & Ishiguro, H. (2018). A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters, 3*(4), 3757–3764. https://doi.org/10.1109/LRA.2018.2856281

Ishii, R., Katayama, T., Higashinaka, R., & Tomita, J. (2018). Generating body motions using spoken language in dialogue. *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 87–92. https://doi.org/10.1145/3267851.3267866

Janhunen, K. (2012). A comparison of likert-type rating and visually-aided rating in a simple moral judgment experiment. *Quality & Quantity, 46*, 1471–1477. https://doi.org/10.1007/s11135-011-9461-x

Janssens, R., Wolfert, P., Demeester, T., & Belpaeme, T. (2022). 'cool glasses, where did you get them?" generating visually grounded conversation starters for human-robot dialogue. *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 821–825. https://doi.org/10.1109/HRI53351.2022.9889489

Jonell, P., Kucherenko, T., Henter, G. E., & Beskow, J. (2020). Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3383652.3423911

Jonell, P., Kucherenko, T., Torre, I., & Beskow, J. (2020). Can we trust online crowdworkers? comparing online and offline participants in a preference test of virtual agents. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3383652.3423860

Jonell, P., Yoon, Y., Wolfert, P., Kucherenko, T., & Henter, G. E. (2021). Hemvip: Human evaluation of multiple videos in parallel. *Proceedings of the*

2021 International Conference on Multimodal Interaction, 707–711. https://doi.org/10.1145/3462244.3479957

Jordan, P. W. (2020). *An introduction to usability*. CRC Press.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1), 81–93. https://doi.org/10/ch8zq6

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.). Charles Griffin & Co.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance, 207–228. https://doi.org/10.1515/9783110813098.207

Kim, H., Ha, Y., Bien, Z., & Park, K. (2012). Gesture encoding and reproduction for human-robot interaction in text-to-gesture systems. *Industrial Robot, 39*(6), 551–563. https://doi.org/10.1108/01439911211268705

Kim, J., Kim, W. H., Lee, W. H., Seo, J.-H., Chung, M. J., & Kwon, D.-S. (2012). Automated robot speech gesture generation system based on dialog sentence punctuation mark extraction. *2012 IEEE/SICE International Symposium on System Integration (SII)*, 645–647. https://doi.org/10.1109/SII.2012.6427293

Kipp, M., & Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–8. https://doi.org/10.1109/ACII.2009.5349544

Kipp, M., Neff, M., Kipp, K. H., & Albrecht, I. (2007). Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent virtual agents. iva 2007. lecture notes in computer science* (Vol. 4722). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74997-4_2

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Wadsworth, Cengage Learning.

Kohavi, R., & Longbotham, R. (2017). Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining, 7*(8), 922–929.

Kong, A. P. H., Law, S. P., Kwan, C. C. Y., & et al. (2015). A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage). *Journal of Nonverbal Behavior, 39*, 93–111. https://doi.org/10.1007/s10919-014-0200-6

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., & Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier (Eds.), *Intelligent virtual agents* (pp. 205–217). Springer Berlin Heidelberg.

Kopp, S., Tepper, P., & Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. *Proceedings of*

the 6th International Conference on Multimodal Interfaces, 97–104. https://doi.org/10.1145/1027933.1027952

Korzun, V., Dimov, I., & Zharkov, A. (2020). The finemotion entry to the genea challenge 2020. *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088609

Kucherenko, T., Hasegawa, D., Henter, G. E., Kaneko, N., & Kjellström, H. (2019). Analyzing input and output representations for speech-driven gesture generation. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 97–104. https://doi.org/10.1145/3308532.3329472

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., & Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. *Proceedings of the 2020 International Conference on Multimodal Interaction*, 242–250. https://doi.org/10.1145/3382507.3418815

Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., & Henter, G. E. (2021). A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 11–21. https://doi.org/10.1145/3397481.3450692

Kucherenko, T., Jonell, P., Yoon, Y., Wolfert, P., Yumak, Z., & Henter, G. (2021). Genea workshop 2021: The 2nd workshop on generation and evaluation of non-verbal behaviour for embodied agents. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 872–873. https://doi.org/10.1145/3462244.3480983

Kucherenko, T., Wolfert, P., Yoon, Y., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2024). Evaluating gesture generation in a large-scale open challenge: The genea challenge 2022 [Just Accepted]. *ACM Trans. Graph.* https://doi.org/10.1145/3656374

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10/dg3k

Lapakko, D. (2015). Communication is 93% nonverbal: An urban legend proliferates. *Communication and Theater Association of Minnesota Journal, 34*(1). https://doi.org/10.56816/2471-0032.1000

Le, Q., Huang, J., & Pelachaud, C. (2012). A common gesture and speech production framework for virtual and physical agents. *ACM international conference on multimodal interaction*.

Le, Q. A., & Pelachaud, C. (2012). Evaluating an expressive gesture model for a humanoid robot: Experimental results. *Submitted to 8th ACM/IEEE International Conference on Human-Robot Interaction*.

Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S. S., & Sheikh, Y. (2019). Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 763–772.

Lemaignan, S., Garcia, F., Jacq, A., & Dillenbourg, P. (2016). From real-time attention assessment to "with-me-ness" in human-robot interaction. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 157–164. https://doi.org/10.1109/HRI.2016.7451747

Levine, S., Krähenbühl, P., Thrun, S., & Koltun, V. (2010). Gesture controllers. In *Acm siggraph 2010 papers*. Association for Computing Machinery. https://doi.org/10.1145/1833349.1778861

Levine, S., Theobalt, C., & Koltun, V. (2009). Real-time prosody-driven synthesis of body language. In *Acm siggraph asia 2009 papers*. Association for Computing Machinery. https://doi.org/10.1145/1661412.1618518

Liang, W., Zou, J., & Yu, Z. (2020). Beyond user self-reported Likert scale ratings: A comparison model for automatic dialog evaluation (D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault, Eds.), 1363–1374. https://doi.org/10.18653/v1/2020.acl-main.126

Lu, J., Liu, T., Xu, S., & Shimodaira, H. (2020). Double-dcccae: Estimation of sequential body motion using wave-form - allthesmooth. *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088376

Lucca, K., & Wilbourn, M. P. (2018). Communicating to learn: Infants' pointing gestures result in optimal learning. *Child Development*, *89*, 941–960. https://doi.org/10.1111/cdev.12707

Lucero, C., Zaharchuk, H., & Casasanto, D. (2014). Beat gestures facilitate speech production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36).

Lydersen, S., Fagerland, M. W., & Laake, P. (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine*, *28*, 1159–1175. https://doi.org/10.1002/sim.3531

Maatman, R. M., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Intelligent virtual agents. iva 2005. lecture notes in computer science* (pp. 25–36, Vol. 3661). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11550617_3

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., & Shapiro, A. (2013). Virtual character performance from speech. *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 25–35. https://doi.org/10.1145/2485895.2485900

Martinez, H., Yannakakis, G., & Hallam, J. (2014). Don't classify ratings of affect; rank them! *IEEE Transactions on Affective Computing*, *3045*(100), 1–1. https://doi.org/10/f6pnzt

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, *8*, 18–25.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10/br5ffs

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

McNeill, D. (2019). *Gesture and thought*. University of Chicago press.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, *6*(1), 109–114. https://doi.org/10.1037/h0024532

Mehta, S., Wang, S., Alexanderson, S., Beskow, J., Székely, É., & Henter, G. E. (2023). Diff-ttsg: Denoising probabilistic integrated speech and gesture synthesis, 150–156. https://doi.org/10.21437/SSW.2023-24

Mlakar, I., Kačič, Z., & Rojc, M. (2013). Tts-driven synthetic behaviour-generation model for artificial bodies. *International Journal of Advanced Robotic Systems*, *10*(10), 344. https://doi.org/10.5772/56870

Mlakar, I., Kačič, Z., & Rojc, M. (2014). Describing and animating complex communicative verbal and nonverbal behavior using eva-framework. *Applied Artificial Intelligence*, *28*(5), 470–503. https://doi.org/10.1080/08839514.2014.905819

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *International Journal of Surgery*, *8*(5), 336–341. https://doi.org/https://doi.org/10.1016/j.ijsu.2010.02.007

Morasso, P. (1981). Spatial control of arm movements. *Experimental Brain Research*, *42*, 223–227. https://doi.org/10.1007/BF00236911

Mueser, K. T., Grau, B. W., Sussman, S., & Rosen, A. J. (1984). You're only as pretty as you feel: Facial expression as a determinant of physical attractiveness. *Journal of Personality and Social Psychology*, *46*(2), 469–478. https://doi.org/10.1037/0022-3514.46.2.469

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. *Proceedings of the 15th International Conference on Intelligent User Interfaces*, 139–148. https://doi.org/10.1145/1719970.1719990

Neerincx, A., Leven, J., Wolfert, P., & de Graaf, M. M. (2023). The effect of simple emotional gesturing in a socially assistive robot on child's engagement at a group vaccination day. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 162–171. https://doi.org/10.1145/3568162.3576960

Neff, M., Kipp, M., Albrecht, I., & Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, *27*(1). https://doi.org/10.1145/1330511.1330516

Ng-Thow-Hing, V., Luo, P., & Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4617–4624. https://doi.org/10.1109/IROS.2010.5654322

Nikulin, M. S. (2001). Hellinger distance. In *Encyclopedia of mathematics*. Springer. http://encyclopediaofmath.org/index.php?title=Hellinger_distance

Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., & Neff, M. (2023). A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum*, *42*, 569–596. https://doi.org/10.1111/cgf.14776

Oetringer, D., Wolfert, P., Deschuyteneer, J., Thill, S., & Belpaeme, T. (2021). Communicative function of eye blinks of virtual avatars may not translate onto physical platforms. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 94–98. https://doi.org/10.1145/3434074.3447136

Ondras, J., Celiktutan, O., Bremner, P., & Gunes, H. (2021). Audio-driven robot upper-body motion synthesis. *IEEE Transactions on Cybernetics, 51*(11), 5445–5454. https://doi.org/10.1109/TCYB.2020.2966730

Pang, K., Komura, T., Joo, H., & Shiratori, T. (2020). Cgvu: Semantics-guided 3d body gesture synthesis. *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4090879

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research, 22*(1), 2617–2680.

Pérez-Mayos, L., Farrús, M., & Adell, J. (2020). Part-of-speech and prosody-based approaches for robot speech and gesture synchronization. *Journal of Intelligent & Robotic Systems, 99*, 277–287. https://doi.org/10.1007/s10846-019-01100-3

Phelps, A. S., Naeger, D. M., Courtier, J. L., Lambert, J. W., Marcovici, P. A., Villanueva-Meyer, J. E., & MacKenzie, J. D. (2015). Pairwise comparison versus likert scale for biomedical image assessment. *American Journal of Roentgenology, 204*(1), 8–14. https://doi.org/10.2214/AJR.14.13022

Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent virtual agents. iva 2010. lecture notes in computer science* (pp. 146–158, Vol. 6356). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15892-6_16

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A. Y., et al. (2009). Ros: An open-source robot operating system. *ICRA workshop on open source software, 3*(3.2), 5.

Rebol, M., Gütl, C., & Pietroszek, K. (2021). Passing a non-verbal turing test: Evaluating gesture animations generated from speech. *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 573–581. https://doi.org/10.1109/VR50410.2021.00082

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Rojc, M., Mlakar, I., & Kačič, Z. (2017). The tts-driven affective embodied conversational agent eva, based on a novel conversational-behavior generation algorithm. *Engineering Applications of Artificial Intelligence, 57*, 80–104. https://doi.org/https://doi.org/10.1016/j.engappai.2016.10.006

Sadoughi, N., & Busso, C. (2019). Speech-driven animation with meaningful behaviors. *Speech Commun., 110*, 90–100. https://doi.org/10.1016/j.specom.2019.04.005

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics, 5*, 313–323. https://doi.org/10.1007/s12369-013-0196-9

Salem, M., Kopp, S., & Joublin, F. (2013). Closing the loop: Towards tightly synchronized robot gesture and speech. In G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social robotics. icsr 2013. lecture notes in computer science* (Vol. 8239). Springer, Cham. https://doi.org/10.1007/978-3-319-02675-6_38

Salem, M., Kopp, S., Wachsmuth, I., & et al. (2012). Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics, 4*, 201–217. https://doi.org/10.1007/s12369-011-0124-9

Salem, M., Rohlfing, K., Kopp, S., & Joublin, F. (2011). A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. *2011 RO-MAN*, 247–252. https://doi.org/10.1109/ROMAN.2011.6005285

Saunderson, S., & Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics, 11*, 575–608. https://doi.org/10.1007/s12369-019-00523-0

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem., 36*(8), 1627–1639. https://doi.org/h10.1021/ac60214a047

Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., & Herre, J. (2018). Webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software, 6*(1).

Schrum, M. L., Johnson, M., Ghuy, M., & Gombolay, M. C. (2020). Four years in review: Statistical practices of likert scales in human-robot interaction studies. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 43–52. https://doi.org/10.1145/3371382.3380739

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc., 63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4799–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Shimazu, A., Hieida, C., Nagai, T., Nakamura, T., Takeda, Y., Hara, T., Nakagawa, O., & Maeda, T. (2018). Generation of gestures during presentation for humanoid robots. *2018 27th IEEE International Symposium*

on Robot and Human Interactive Communication (RO-MAN), 961–968. https://doi.org/10.1109/ROMAN.2018.8525621

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10/gfxv4h

Straube, B., Green, A., Bromberger, B., & Kircher, T. (2011). The differentiation of iconic and metaphoric gestures: Common and unique integration processes. *Human Brain Mapping, 32*, 520–533. https://doi.org/10.1002/hbm.21041

Sung, Y.-T., & Wu, J.-S. (2018). The visual analogue scale for rating, ranking and paired-comparison (vas-rrp): A new technique for psychological measurement. *Behavior Research, 50*, 1694–1715. https://doi.org/10.3758/s13428-018-1041-8

Takeuchi, K., Hasegawa, D., Shirakawa, S., Kaneko, N., Sakuta, H., & Sumi, K. (2017). Speech-to-gesture generation: A challenge in deep learning approach with bi-directional lstm. *Proceedings of the 5th International Conference on Human Agent Interaction*, 365–369. https://doi.org/10.1145/3125739.3132594

Thangthai, A., Thangthai, K., Namsanit, A., Thatphithakkul, S., & Saychum, S. (2020). The nectec gesture generation system entry to the genea challenge 2020. *Proc. GENEA Workshop*. https://doi.org/10.5281/zenodo.4088629

Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. In B. Raj & J. Koerts (Eds.), *Henri theil's contributions to economics and econometrics: Econometric theory and methodology* (pp. 345–381). Springer. https://doi.org/10.1007/978-94-011-2546-8_20

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Sage.

Union, E. B. (2020). Loudness normalisation and permitted maximum level of audio signals. https://tech.ebu.ch/docs/r/r128.pdf

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics, 61*, 89–101. https://doi.org/10.1007/BF00204593

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209–232. https://doi.org/https://doi.org/10.1016/j.specom.2013.09.008

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://doi.org/https://doi.org/10.1016/j.ijresmar.2010.02.004

Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the godspeed questionnaire series. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 381–388. https://doi.org/10.1109/ROMAN.2015.7333568

Wester, M., Wu, Z., & Yamagishi, J. (2016). Analysis of the Voice Conversion Challenge 2016 evaluation results. *Proc. Interspeech*, 1637–1641.

Wolfert, P., De Gersem, L., Janssens, R., & Belpaeme, T. (2024). Multi-modal language learning: Explorations on learning japanese vocabulary. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 1129–1133. https://doi.org/10.1145/3610978.3640685

Wolfert, P., Deschuyteneer, J., Oetringer, D., Robinson, N., & Belpaeme, T. (2020). Security risks of social robots used to persuade and manipulate: A proof of concept study. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 523–525. https://doi.org/10.1145/3371382.3378341

Wolfert, P., Girard, J. M., Kucherenko, T., & Belpaeme, T. (2021). To rate or not to rate: Investigating evaluation methods for generated co-speech gestures. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 494–502. https://doi.org/10.1145/3462244.3479889

Wolfert, P., Henter, G. E., & Belpaeme, T. (2023). "am i listening?", evaluating the quality of generated data-driven listening motion. *Companion Publication of the 25th International Conference on Multimodal Interaction*, 6–10. https://doi.org/10.1145/3610661.3617160

Wolfert, P., Henter, G. E., & Belpaeme, T. (2024). Exploring the effectiveness of evaluation practices for computer-generated nonverbal behaviour. *Applied Sciences*, *14*(4). https://doi.org/10.3390/app14041460

Wolfert, P., Kucherenko, T., Kjellström, H., & Belpaeme, T. (2019). Should beat gestures be learned or designed?: A benchmarking user study. *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*.

Wolfert, P., Kucherenko, T., Viegas, C., Yumak, Z., Yoon, Y., & Henter, G. E. (2022). Genea workshop 2022: The 3rd workshop on generation and evaluation of non-verbal behaviour for embodied agents. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 799–800. https://doi.org/10.1145/3536221.3564027

Wolfert, P., Robinson, N., & Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, *52*(3), 379–389. https://doi.org/10.1109/THMS.2022.3149173

Xu, Y., Pelachaud, C., & Marsella, S. (2014). In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Intelligent virtual agents. iva 2014. lecture notes in computer science* (Vol. 8637). Springer, Cham. https://doi.org/10.1007/978-3-319-09767-1_58

Yannakakis, G., Cowie, R., & Busso, C. (2021). The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, *12*(1), 16–35. https://doi.org/10.1109/TAFFC.2018.2879512

Yannakakis, G., & Martínez, H. P. (2015). Grounding truth via ordinal annotation. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 574–580. https://doi.org/10/gjp74q

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., & Lee, G. (2020). Speech gesture generation from the trimodal context of text,

audio, and speaker identity. *ACM Trans. Graph.*, *39*(6). https://doi.org/10.1145/3414685.3417838

Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., & Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. *2019 International Conference on Robotics and Automation (ICRA)*, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2022). The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. *Proceedings of the 2022 International Conference on Multimodal Interaction*, 736–747. https://doi.org/10.1145/3536221.3558058