Minimal Latency Speech-Driven Gesture Generation for Continuous Interaction in Social XR

1st Niklas Krome
Faculty of Technology
Bielefeld University
Bielefeld, Germany
nkrome@techfak.uni-bielefeld.de

2nd Stefan Kopp
Faculty of Technology
Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

***Abstract*—Social XR applications usually require advanced tracking equipment to control one's own avatar. We explore if AI-based co-speech gesture generation techniques can be employed to compensate for the lack of tracking hardware that many users face. One main challenge is to achieve convincing behavior quality without introducing too much latency. Previous work has shown that both depend – in opposite ways – on the length of the audio chunk the gestures are generated from, and that gesture quality of existing models declines with lower chunk sizes while still not reaching sufficiently low latency to enable fluent interaction. In this paper we present an approach that is able to generate continuous gesture trajectories frame by frame, minimizing latency and yielding delays well below buffer sizes of voice communication systems or video calls. A project page with videos of the generated gestures is available at https://nkrome.github.io/FrameCAGE.html.**

***Index Terms*—extended reality, social interaction, animation, machine learning, gesture generation**

I. Introduction

Social VR applications provide an immersive and life-like way of interacting over long distances. They do however suffer from an accessibility issue, as they require expensive motion tracking hardware. Modern generative AI methods can provide a way of alleviating this issue by generating synthetic avatar behavior from audio information [1]–[11], which is available on all devices. This would enable eXtended Reality (XR) applications where different users with different devices can join a shared virtual environment as illustrated in Fig. (1).

We pursue this vision with a focus on augmenting avatar behavior through AI-based co-speech gesture generation. Recent work [29] investigated the use of a co-speech gesture generation model to infer full-body motion from a continuous stream of live audio input in an online interaction. Current models for co-speech gesture generation use deep neural networks to correlate speech audio with appropriate full-body motion [1]–[11]. Subjective evaluation shows that these models produce smooth and indistinguishably human-like motion, while still lacking a good relation to speech content [27]. However, human-likeness of motion is more important for user immersion than communicative function, making existing AI models suitable for integration into Social XR applications. This integration however is difficult, because these models are not tailored for online interaction. They are usually not optimized in terms of inference speed and exploit auditory information of the whole utterance to match the rhythm of the speech. This information, however, is not available in a real-time scenario.

In previous work [29] we explored an incremental gesture generation approach, by providing an AI model with a constant stream of live audio information. This information was split into chunks of varying lengths from several seconds down to half a second. By changing the chunk size, we restricted the amount of information the model got access to and evaluated the resulting behavior in a user study, showing a steep decline in subjective ratings for gestures generated from lower chunk sizes. The length of the initial increment represents the major portion of the resulting latency, when using the system in online interaction and even the lowest chunk size of 0.5 seconds caused too much delay to enable a fluent conversation [28], [30], [31].

In this paper we present a substantial extension of this work, by presenting a model that allows for a minimal latency of down to a single frame, without sacrificing motion quality. In the following, we briefly discuss related work regarding co-speech gesture generation, leading to the basic ideas behind an incremental approach to handle live audio. We identify the reasons for a decline in gesture quality when using smaller audio chunks, and we present a new pipeline that minimizes the chunk size while retaining the quality of the baseline system. We evaluate this approach with a human rater study in direct comparison to the findings with larger chunk sizes in the previous work. Finally, we discuss the implications of our findings and provide an outlook on ways how to enhance Social XR using AI.

II. Related Work

Co-speech gesture generation describes the process of inferring non-verbal behavior from a given utterance. While early systems relied on a set of hand-crafted rules [24], [25], the current state-of-the-art is dominated by different neural network architectures [1]–[21], directly mapping speech audio, text, or both to full-body motion. Rule-based systems generated gestures that contained lots of information and were aimed at increasing the communicative efficacy of virtual agents. Data-driven systems, on the other hand, aim for behavioral realism

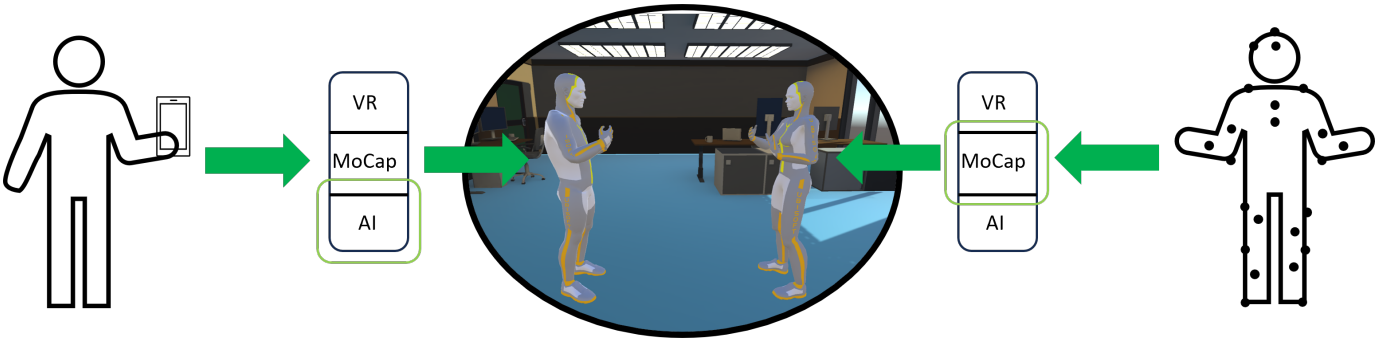

Fig. 1. An overview of the envisioned environment with motion captured and not captured users participating in a shared space. Depending on the input device, a different behavior generation module activates to transfer or generate full-body motion.

in terms of human-likeness of the motion. This is achieved by analysing frequencies in speech audio, to accurately produce beat gestures, matching the speech rhythm. The resulting behavior is perceived as natural and plausible by human raters [27]. Some systems exploit text information to create a link to the speech content and produce more meaningful gestures. However, this is still an active research direction as even the most sophisticated models produce gestures that are rated significantly less appropriate than ground truth motion.

Most current systems rely on sequential techniques like auto-regression and recurrent architectures to achieve a smooth motion sequence; other systems even smooth the resulting gestures afterwards to eliminate jerkiness. Some approaches produce the output gesture sequence all at once, whereas others create it frame by frame allowing for generating a non-predefined sequence length. Recent work has proposed to use diffusion models, but the performance is not significantly better and these systems are extremely slow [32]–[34]. Different ways have been investigated to control the gestures a model produces. This control may consist of manipulating principle components of the learned gesture space or directly affecting gesture radius, handedness, or height [1], [2], [14]. Gestures can also be controlled via style presets describing a particular way of gesticulating, attributed to a specific person, emotion, or situation [2], [15]. For this, the model either has to be fine-tuned on a data set of the desired speaker [15], or a style embedding is calculated and fed to the model at runtime [2]. An extensive overview over the different models is given by Nyatsanga et al. [26].

The model that was used in our previous work [29] is the ZeroEGGS model by Ghorbani et al. [2]. This model was chosen as its performance outperforms other audio based models and as it fulfills the requirements for integration into an incremental real-time generation pipeline. The model takes speech audio as input, as well as a style example. It then produces gestures in an Encoder-Decoder fashion, by calculating speech and style embeddings and feeding the concatenated vector to a recurrent decoder, which then updates the character pose frame by frame. The accompanying data set includes 19 different styles, ranging from emotions like happy, sad, or angry, to discourse contexts like agreement and disagreement. The style embeddings can also be combined via linear interpolation, covering an extensive space of gesturing styles. Apart from style control, ZeroEGGS also allows for the specification of a starting pose, when generating a gesture sequence which is especially useful for generating partial gestures and stitching them together to form a continuous sequence.

III. Model

Building on the ZeroEGGS model, we have proposed an incremental processing approach that slices the incoming audio stream into chunks of predefined length, generates gesture segments for a chunk, and ensures continuity across chunk borders [29]. The resulting gestures were incrementally combined and animated on a virtual avatar in Unity. This approach was evaluated for chunk sizes ranging from 5 seconds down to 0.5 seconds, showing a steep decrease in subjectively perceived gesture quality for smaller chunk sizes. One reason for this is that shorter audio sequences result in a different frequency spectrum, over-representing lower frequencies in the audio data, which in turn leads to a faster gesture rhythm that does not accurately match slow speech sections, which in turn leads to lower ratings in terms of temporal synchrony and appropriateness. Additionally, when working with partial gestures derived from smaller audio chunks, the continuity of the final motion sequence has to be guaranteed by enforcing the starting point of each chunk to be equal or similar to the last frame of the previous chunk. To prevent jumps, chunks are then blended together slightly which smooths out the gestures and damps the motion, especially for very small chunk sizes.

To overcome these limitations, we extended the previous model with a long-term memory that preserves and carries over temporal information across successive increments. To that end we introduce a latent vector, containing pose and style information, that is updated with each increment. It effectively initializes and saves the cell state of the recurrent decoder before and after each inference. This allows the model to utilize the same amount of long-term information as if the input sequence was given all at once. This way, we can in principle minimize the resulting latency by reducing the chunk size even down to the equivalent of a single frame. Further,

we also do not have to interpolate between animations as we are working with single keyframes.

In the new model, we keep the general design of the generation pipeline, divided into audio recording, speech processing, gesture generation and visualization. We start by pre-calculating a style embedding and initialize a starting pose as well as a hidden decoder state before the generation process. We then read an audio chunk equivalent to a single frame from the input audio stream and calculate a singular speech embedding. We then predict a body pose and the next decoder state, based on the speech embedding, the style embedding and the current decoder state. The body pose is represented as a sequence of joint angles, which are passed to our visualization environment. In Unity, the joint angles are parsed and used to create a keyframe for the target avatar, which is played immediately as it becomes available. After a frame has been generated the decoder state is updated and the next audio frame is processed to generate the next pose. The new input then consist of a new speech embedding, a new style embedding and the updated decoder state from the previous inference. An overview of the architecture is shown in Fig. (2).

In the previous model the lowest possible chunk size was 0.5 seconds, which led to an overall delay between audio input and animation playback of around 900ms. However, the recommended latency for applications like Zoom is around 150ms or lower [28], so when accounting for inference time and network latency, the chunk size needs to undercut that. A single frame of the avatar animation is equivalent to 16.7ms at a frame rate of 60 fps and can be generated within 4ms, which is well below the required latency, enabling the continuous generation in online interaction. The new model presented here allows for any number of frames to be generated per inference step, down to a single frame. More frames lead to a longer audio sequence and a larger speech embedding vector. The frequency information is extracted by calculating a mel-spectrogram with a filter length of 800 samples and a hop length of 200 samples. Longer input sequences therefore increase the amount of information the model can gather, but also increase the resulting latency. As the generation of a single gesture frame takes about a fourth as long as playing the frame during visualization, it is possible to increase the audio window a single frame is based on by up to three times, before running into continuity constraints in the visualization, because frames are not produced in time.

IV. Evaluation

Our approach for frame-wise gesture generation allowed for the animation of continuous full-body motion on an avatar. From visual inspection, we could not tell a difference in quality when providing shorter or longer input audio sequences. To more thoroughly evaluate the performance and compare it against the previous incremental system, we conducted a human rater study analogous to [29], keeping the baseline condition as a comparative measure and adding conditions with different frame counts well below the previously lowest chunk size.

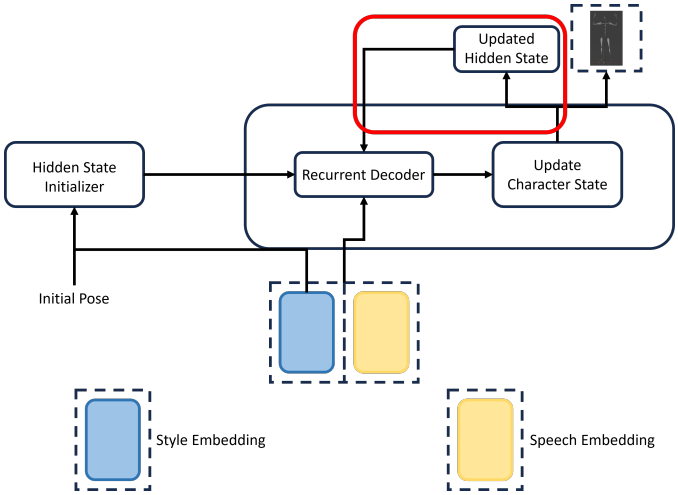

Fig. 2. An overview of the architecture for generating a body pose from a single speech embedding, adapted from [2] and extended by the hidden state update (outlined in red).

A. Procedure

To investigate the effects of generating different amounts of frames at once, we recorded videos of the different conditions animating gestures in synchrony with the corresponding speech audio and had them rated by human participants in an online study. We utilized a playback process to simulate a real-time audio recording with controllable input, which reads a sound file and sends a specific amount of audio samples to the generation process, waiting in between each increment for the corresponding duration as if the audio was becoming available in real-time. We used the 5 styles from the previous study, which originated from the source data set by Ghorbani et al. [2]. These were ”Angry”, ”Neutral”, ”Relaxed”, ”Sad” and ”Speech”. Each audio file was processed in chunks of 20, 10 and 5 frames, as well as directly frame-by-frame. We chose these specific sets to cover the lowest possible frame count of 1, as well as an amount of samples that exceeds the length of the underlying audio filter at 5 frames, or 1335 samples. Also, 10 frames correspond to about 160ms of audio, which is about the maximum recommended latency for online interaction and we tested 20 frames to see if there was an improvement when increasing latency.

The study was performed via Prolific [23], using a Questionnaire by SoSciSurvey [22]. 50 participants, 25 male and 25 female, all native speakers of English, took part in the study. In a within-subject design, they were each shown every video and were instructed to watch the videos in their entirety and then evaluate the behavior of the avatar by answering 7 questions on a 7 point Likert scale. The questions were aimed to capture the most common metrics for evaluating generated gestures. These were ”Human-likeness”, ”Smoothness”, ”Content Matching” and ”Beat Matching”. They should give us an idea if the generated gestures contain motion untypical for humans, or overly erratic behavior, as well as if the produced gestures match the content and rhythm of the corresponding speech

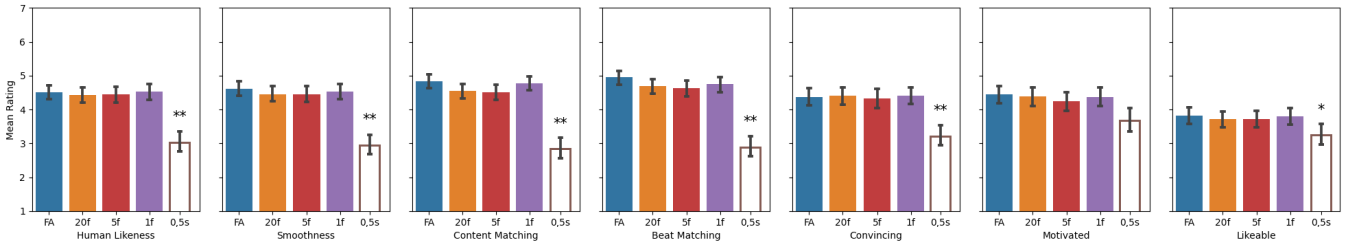

Fig. 3. Results of the evaluation study, comparing gestures produced by processing 20, 5 or 1 frame of input audio at a time, as well as the control condition using full audio information (FA). The hollow bar illustrates the ratings of the lowest latency condition (0.5s chunks, or 30 frames), when using the previous system presented in [29]. Mean subjective ratings are given with confidence intervals and significance levels relative to ZeroEGGS, when given the full audio, according to Dunn test with bonferroni correction.

segment. We also asked if the speaker seemed "Convincing", "Motivated", or "Likeable" to capture the general reception of the avatars. We included attention checks in regular intervals to exclude inattentive raters. Raters that passed all attention checks, but answered the questions within well below 20 seconds were also excluded, because they could not have watched the entire video. 12 participants were excluded this way.

B. Results

The results are given in Figure (3). All conditions received similar ratings between 4 and 5 on all scales, except the "Likeability" scale, which shows ratings slightly below 4 for all conditions. That is, we can confirm the consistent performance of our model, independent of the number of audio frames inputted at once when processing in an incremental manner. In result, our pipeline no longer impacts the gesture quality of the baseline model, when generating gestures incrementally. We have also included a result from our previous study [29], visualized as a hollow bar in Figure (3). This was the average rating for the lowest latency condition when using the old system, generating gestures in 0.5 second chunks, which is equivalent to generating 30 frames with the new system. We are now able to achieve significantly better ratings, while reducing latency well below the prior minimum. However, it is important to note that these results are gathered from presenting participants with videos that were generated offline and using a simulated real-time audio source (as in our previous study [29]). This controlled input led to visually indistinguishable behavior for the different conditions. When testing the pipeline with actual live audio from a microphone, there was a small but noticeable difference in beat synchronicity. We plan on performing an interaction study in a networked environment to evaluate the system on actual live audio. Additionally, some of the participants stated, that the study was boring and tedious, which was not noted on the previous study. This may be due to the extreme similarity of the motion between the conditions. After a while, participants may have felt indifferent about the videos and rated every video similarly without thoroughly investigating the motion. To capture more subtle differences, we plan on comparing different conditions side by side in our next study. This may also help to keep participants engaged, as they no longer have to watch several identical looking motion sequences in short succession.

V. Conclusion

In this paper we have presented an approach to use a generative AI model to create co-speech gestures from live audio to augment Social XR scenarios. We were able to minimize the system's latency while maintaining the resulting gesture quality to match that of the baseline system, by implementing a frame-wise pipeline with a latent state as persistent long-term memory over multiple inferences. We evaluated the subjective perception of the generated behavior and compared it to the previous results. The performance was improved significantly in every respect, allowing the integration into an online interaction scenario without a quality trade-off and opening up multiple avenues for further research. We envision several follow-up studies to investigate the perception of generated gestures in a real interaction, both from an observer standpoint, as well as from the perspective of the user. Also there are ways to provide more agency to the user of a generative system. By including more input modalities, such as capturing facial features and head position via webcam or the front camera of a smartphone, an appropriate style setting could be detected in real-time to adapt the gestures to the user's mood for example.

In sum, AI technology holds a lot of potential to process different input modalities in order to enhance Social XR scenarios by augmenting the environment or supplementing missing information. This opens up new possibilities to increase the user's immersion or sense of agency in VR or XR.

References

[1] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows", Computer Graphics Forum, vol. 39, no. 2, pp. 487–496, May 2020, doi: https://doi.org/10.1111/cgf.13946.

[2] S. Ghorbani, Y. Ferstl, D. Holden, N. F. Troje, and M. Carbonneau, "ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech", Computer Graphics Forum, vol. 42, no. 1, pp. 206–216, Feb. 2023, doi: https://doi.org/10.1111/cgf.14734.

[3] I. Habibie, M. Elgharib, K. Sarkar, A. Abdullah and S. Nyatsanga, "A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech", Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings, Aug. 2022, doi: https://doi.org/10.1145/3528233.3530750.

[4] B. Wu, C. Liu, C. T. Ishi, and H. Ishiguro, "Probabilistic Human-like Gesture Synthesis from Speech using GRU-based WGAN", Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), ACM, Oct. 2021, doi: https://doi.org/10.1145/3461615.3485407.

[5] S. L. Taylor, J. Windle, D. A. Greenwood, and I. Matthews, "Speech-Driven Conversational Agents using Conditional Flow-VAEs", Proceedings of the 18th ACM SIGGRAPH European Conference on Visual Media Production (CVMP '21), ACM, Dec. 2021, doi: https://doi.org/10.1145/3485441.3485647.

[6] M. Rebol, C. Gütl and K. Pietroszek, "Passing a Non-verbal Turing Test: Evaluating Gesture Animations Generated from Speech", 2021 IEEE Virtual Reality and 3D User Interfaces (VR), Lisboa, Portugal, 2021, pp. 573-581, doi: 10.1109/VR50410.2021.00082.

[7] S. Qian, Z. Tu, Y. Zhi, W. Liu and S. Gao, "Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates", in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 11057-11066. doi: 10.1109/ICCV48922.2021.01089

[8] Y. Ferstl, M. Neff, and R. McDonnell, "ExpressGesture: Expressive gesture generation from speech through database matching", Computer Animation and Virtual Worlds, vol. 32, no. 3–4, May 2021, doi: https://doi.org/10.1002/cav.2016.

[9] J. Li, D. Kang, W. Pei, X. Zhe, Y.Zhan, Z. He and L. Bao, "Audio2Gestures: Generating Diverse Gestures From Audio", IEEE Transactions on Visualization and Computer Graphics, pp. 1–15, 2023, doi: https://doi.org/10.1109/tvcg.2023.3276973.

[10] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib and C. Theobalt, "Learning Speech-driven 3D Conversational Gestures from Video", Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA '21), ACM, Sep. 2021, doi: https://doi.org/10.1145/3472306.3478335.

[11] K. J. Saleh, "Hybrid Seq2Seq Architecture for 3D Co-Speech Gesture Generation", Proceedings of the 2022 International Cconference on Multimodal Interaction (ICMI '22), ACM, Nov. 2022, doi: https://doi.org/10.1145/3536221.3558064.

[12] N. Kaneko, Y. Mitsubayashi and G. Mu, "TransGesture: Autoregressive Gesture Generation with RNN-Transducer", Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22), ACM, Nov. 2022, doi: https://doi.org/10.1145/3536221.3558061.

[13] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexanderson, I. Leite and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation", Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), ACM, Oct. 2020, doi: https://doi.org/10.1145/3382507.3418815.

[14] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic Gesticulator", ACM Transactions on Graphics, vol. 41, no. 6, pp. 1–19, Nov. 2022, doi: https://doi.org/10.1145/3550454.3555435.

[15] C. Ahuja, D. W. Lee and L. -P. Morency, "Low-Resource Adaptation for Personalized Co-Speech Gesture Generation", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 20534-20544, doi: 10.1109/CVPR52688.2022.01991.

[16] W. Zhuang, J. Qi, P. Zhang, B. Zhang, and P. Tan, "Text/Speech-Driven Full-Body Animation", Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI '22) Demo Track. pp. 5956-5959. 2022.

[17] Y. Liang, Q. Feng, L. Zhu, L. Hu, P. Pan and Y. Yang, "SEEG: Semantic Energized Co-speech Gesture Generation", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10463-10472, doi: 10.1109/CVPR52688.2022.01022.

[18] M. Fares, C. Pelachaud, and N. Obin, "Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding", Frontiers in Artificial Intelligence, vol. 6, p. 1142997, Jun. 2023, doi: https://doi.org/10.3389/frai.2023.1142997.

[19] C. Zhou, T. Bian, and K. Chen, "GestureMaster: Graph-based Speech-driven Gesture Generation", Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22), Nov. 2022, doi: https://doi.org/10.1145/3536221.3558063.

[20] T. Kucherenko, R. Nagy, M. Neff, H. Kjellström and G. E. Henter, "Multimodal Analysis of the predictability of hand-gesture properties", Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS '22), ACM, pp. 770-779. May 2022.

[21] T. V. T. Nguyen and O. Celiktutan, "Context-aware body gesture generation for social robots", ICRA 2022 Workshop on Prediction and Anticipation Reasoning for Human-Robot Interaction. 2022.

[22] "SoSci Survey - the Professional Solution for Your Online Survey", https://www.soscisurvey.de

[23] "Many Projects, One Interface", Prolific, https://www.prolific.com

[24] S. Kopp, B. Jung, N. Leßmann and I. Wachsmuth, "Max - A multimodal assistant in virtual reality construction", KI - Künstliche Intelligenz, vol. 4, 2003, pp. 11-17.

[25] J. Lee and S. Marsella, "Nonverbal Behavior Generator for Embodied Conversational Agents", Intelligent Virtual Agents, pp. 243–255, 2006, doi: https://doi.org/10.1007/11821830_20.

[26] S. Nyatsanga, T. Kucherenko, C. Ahuja, Henter, Gustav Eje, and M. Neff, "A Comprehensive Review of Data-Driven Co-Speech Gesture Generation", vol. 42, no. 2, pp. 569–596, Jan. 2023, doi: https://doi.org/10.1111/cgf.14776.

[27] Y. Yoon, P. Wolfert, T. Kucherenko, C. Viegas, T. Nikolov, M. Tsakov and G. E. Henter, "The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation", Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22), ACM, Nov. 2022, doi: https://doi.org/10.1145/3536221.3558058.

[28] "Accessing meeting and phone statistics - Zoom Support," support.zoom.com. https://support.zoom.us/hc/en-us/articles/202920719-Accessing-meeting-and-phone-statistics (accessed Oct. 19, 2023).

[29] N. Krome and S. Kopp, "Towards Real-time Co-speech Gesture Generation in Online Interaction in Social XR", Proc. of the 23rd International Conference on Intelligent Virtual Agents (IVA '23), ACM, ed., 2023.

[30] J. Holub, M. Kastner and O. Tomiska, "Delay effect on conversational quality in telecommunication networks: Do we mind?", 2007 Wireless Telecommunications Symposium, Pomona, CA, USA, 2007, pp. 1-4, doi: 10.1109/WTS.2007.4563311.

[31] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications", in IEEE Journal on Selected Areas in Communications, vol. 9, no. 4, pp. 586-593, May 1991, doi: 10.1109/49.81952.

[32] A. Deichler, S. Mehta, S. Alexanderson, and J. Beskow, "Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation", Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23), Oct. 2023, doi: https://doi.org/10.1145/3577190.3616117.

[33] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu and L. Yu, "Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation", in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 10544-10553. doi: 10.1109/CVPR52729.2023.01016

[34] S. Alexanderson, R. Nagy, J. Beskow, and Gustav Eje Henter, "Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models", ACM Transactions on Graphics, vol. 42, no. 4, pp. 1–20, Jul. 2023, doi: https://doi.org/10.1145/3592458.