ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika

ISSN(p): 2338-8323 | ISSN(e): 2459-9638 | Vol. 12 | No. 4 | Halaman 953 - 967
DOI : http://dx.doi.org/10.26760/elkomika.v12i4.953 Oktober 2024

Pengembangan dan Evaluasi Agen Virtual dengan Model Generasi Gestur berbasis Aturan Sederhana

MUHAMMAD FIRDAUS SYAWALUDIN LUBIS, MIKAIL FAUZAN ATHALLAH, CATUR APRIONO

Dept. of Electrical Engineering, Faculty of Engineering Universitas Indonesia, Depok, Indonesia
Email: m.firdaus04@ui.ac.id

Received 1 September 2024 **|** *Revised* 30 September 2024 **|** *Accepted* 30 Oktober 2024

ABSTRAK

Studi pengembangan gestur sebelumnya telah menyoroti manfaat pendekatan berbasis deep learning untuk menghasilkan gerakan yang mirip manusia, namun, pendekatan tersebut memerlukan dastaset besar dan komputasi intensif. Model milik penulis membedakan antara dialog pendek dan panjang, menghasilkan gerakan spesifik konteks untuk percapakan pendek (salam, perpisahan, persetujuan/tidak setuju) dan gerakan berbasis emosi untuk dialog yang lebih panjang (netral, bahagia, agresif). Penulis membandingkan kinerja sistem dengan ground truth gestures, random gestures, dan idling gestures, menggunakan metrik dari GENEA Challenge. Pendekatan ini bertujuan untuk memberikan alternatif yang lebih efisien dibandingkan model deep learning. Temuan penulis diharapkan dapat berkontribusi pada pengembangan generasi gestur yang menarik meningkatkan pemahaman pengguna dalam interaksi manusia-komputer.

***Kata kunci**: Generasi Gestur, Interaksi Manusia-Komputer, Ukuran Dialog*

ABSTRACT

While previous gesture generation studies have highlighted the benefits of deep learning-based approaches for generating human-like gestures, these often require large datasets and intensive computation. Our model differentiates between short and long dialogues, generating context-specific gestures for short exchanges (e.g., greetings, farewells, agreement/disagreement) and emotion-based gestures for longer dialogues (neutral, happy, aggressive). We compare the system's performance against ground truth gestures, random gestures, and idling gestures using metrics from the GENEA Challenge. This approach aims to provide a more efficient alternative to deep learning models. Our findings are expected to contribute to the development of more engaging, responsive virtual assistants, improving user comprehension in human-computer interaction.

***Keywords**: Gesture Generation, Human-Computer Interaction, Dialogue Size*

1. PENDAHULUAN

Komunikasi manusia adalah perpaduan kompleks antara kata-kata yang diucapkan dan isyarat non-verbal seperti ekspresi wajah, bahasa tubuh, dan gerakan **(Arnheim dan McNeill, 1994)** menekankan bahwa gerakan sangat penting dalam menyampaikan makna, emosi, dan niat selama interaksi tatap muka. Dalam ranah interaksi manusia-komputer (HCI), kemampuan untuk menghasilkan dan memahami gerakan ini menjadi kunci untuk menciptakan antarmuka yang intuitif dan alami, sebagaimana dijelaskan oleh **(Kopp dan Wachsmuth, 2004)**.

Generasi gestur, yaitu produksi otomatis gerakan oleh sistem komputasi, memungkinkan agen virtual berkomunikasi melalui cara non-verbal. Proses ini melibatkan penerjemahan informasi semantik atau input menjadi gerakan fisik yang dapat dilihat oleh manusia. Dengan memasukkan pembuatan gerakan ke dalam sistem HCI (*Human Computer Interaction*), penulis menjembatani kesenjangan antara cara komunikasi manusia dan komputer, yang berujung pada interaksi yang lebih efektif dan menarik, seperti yang diungkapkan oleh **(Cassell, 2001)**.

Meskipun generasi gestur telah dipelajari secara luas, dampak panjang dialog terhadap persepsi gestur masih belum banyak dieksplorasi. Perilaku manusia sering berubah berdasarkan panjang dan kompleksitas kalimat, yang penting untuk pembuatan gerakan alami dalam asisten virtual. Studi menunjukkan bahwa selama kalimat panjang atau monolog, individu lebih cenderung menggunakan gerakan tubuh yang ekspresif, menggunakan gerakan untuk mendukung isi dan nada emosional dari pidato mereka. Sebaliknya, kalimat yang lebih pendek biasanya disertai dengan gerakan yang lebih sedikit, karena membutuhkan beban kognitif dan penguatan ekspresif yang lebih rendah. Temuan ini mencerminkan perilaku manusia baik dalam percakapan sehari-hari maupun dalam berbicara di depan umum, di mana pembicara sering mengandalkan gerakan selama penjelasan panjang untuk mempertahankan keterlibatan dan kejelasan **(Tipper, dkk, 2015)**.

Penelitian penulis secara khusus membedakan antara dialog pendek dan panjang dalam generasi gestur. Penulis memperkenalkan model pembuatan gerakan berbasis aturan yang inovatif untuk asisten virtual, yang dirancang untuk menangani dialog pendek dan panjang. **(Hostetter dan Alibali, 2008)** membahas mekanisme kognitif di balik integrasi gestur dan bahasa dalam interaksi singkat, yang mendukung pendekatan penulis untuk dialog pendek, termasuk gestur yang spesifik secara konteks seperti salam, perpisahan, dan setuju/tidak setuju. Dalam dialog panjang, pembicara cenderung lebih mengandalkan gestur berbasis emosi (netral, bahagia, agresif) untuk komunikasi non-verbal yang berkelanjutan, yang sejalan dengan temuan **(Krämer, dkk, 2015)**. Oleh karenanya, dalam dialog panjang, penulis fokus pada gerakan berbasis emosi, yang dikategorikan sebagai netral, bahagia, atau agresif, karena pembicara sering beralih ke penggunaan bahasa tubuh yang lebih dinamis untuk menekankan keadaan emosional mereka **(Kendon, 1980)**.

Untuk mengevaluasi efektivitas model yang digunakan, penulis melakukan analisis perbandingan dengan *ground truth*, gerakan acak, dan gerakan diam. Menggunakan metrik yang terinspirasi oleh GENEA Challenge (*Generation and Evaluation of Non-verbal Behaviour for Embodied Agents*), penulis menilai kemiripan dengan manusia dan kesesuaian dalam dialog pendek dan panjang. Eksperimen penulis, melibatkan 43 peserta yang menonton video rekaman asisten virtual dalam peran sebagai pengajar, memberikan wawasan tentang kinerja sistem berbasis aturan milik penulis.

Penelitian ini bertujuan untuk memajukan bidang HCI dan memberikan kontribusi signifikan terhadap pengembangan asisten virtual yang lebih menarik dan realistis. Temuan penulis

menawarkan implikasi penting untuk meningkatkan interaksi manusia-komputer dan mengembangkan asisten virtual yang lebih responsif di masa depan, terutama dengan mempertimbangkan faktor penting dari panjang dialog dalam pembuatan gerakan.

2. METODE PENELITIAN

2.1 Agen *Virtual*

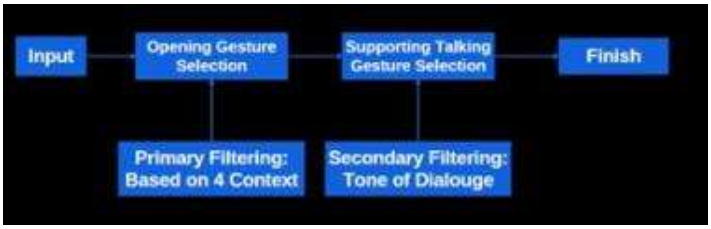
Penulis mengembangkan agen *virtual* ini bernama Kemala. Karakter yang didorong oleh kecerdasan buatan ini dirancang dengan harapan untuk menjembatani komunikasi antara sistem digital dan pengguna manusia. Model Kemala dikembangkan menggunakan VROID Studio, sebuah alat pembuatan karakter 3D yang kuat, dan diintegrasikan ke dalam *Unreal Engine* 5.3. Dalam artikel yang dibuat oleh **(Agarwal , 2023)** yang mencatat perkembangan *Unreal Engine*, perangkat lunak ini adalah alat untuk pengembangan *game*, visualisasi arsitektur dan otomotif, produksi konten film dan televisi, serta aplikasi *real-time* lainnya. Engine ini dikenal karena kemampuannya untuk menghasilkan visual dengan kualitas tinggi, sehingga sering digunakan dalam industri yang mengutamakan kualitas visual. Kombinasi ini memungkinkan kustomisasi yang mendetail pada penampilan dan ekspresi Kemala, sehingga menghasilkan agen *virtual* yang menarik secara visual dan sangat ekspresif.

Gambar 1. Kemala Agen *Virtual*

Untuk kemampuan percakapan yang canggih, Kemala terintegrasi dengan OpenAI ChatCompletion API, secara khusus menggunakan model GPT-3.5-turbo. ChatCompletion API memanfaatkan arsitektur Transformer. Berdasarkan karya **(Vaswani, dkk, 2017)** arsitektur ini unggul dalam menangani dependensi jangka panjang, maka API ini meningkatkan kemampuan pemrosesan data, memungkinkan pengelolaan urutan panjang dan penanganan data secara paralel dengan efisien. Integrasi ini melengkapi agen *virtual* penulis dengan kemampuan pemrosesan bahasa alami yang maju, memungkinkan percakapan dinamis dan kontekstual yang dapat beradaptasi dengan berbagai input dan skenario pengguna. Selain itu, penulis telah mengintegrasikan ElevenLabs API untuk meningkatkan pengalaman interaksi melalui sintesis suara berkualitas tinggi yang alami. Menurut **(Martin, 2024)** dalam artikel ulasan di Technopedia, ElevenLabs muncul sebagai salah satu alat AI terbaik di pasaran, khususnya dalam bidang teknologi sintesis suara.

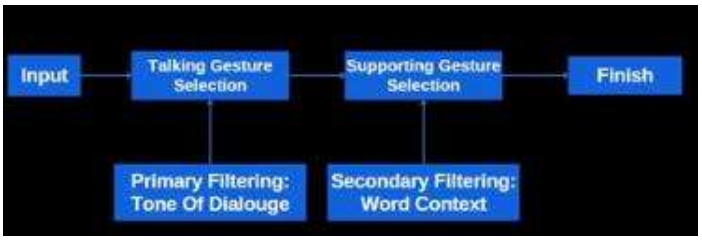
2.2 Generasi Gestur Berbasis Aturan

Sistem generasi gestur penulis dirancang untuk menangani dialog pendek dan panjang, menyesuaikan pendekatannya berdasarkan panjang input. Dalam Mode Dialog Pendek, ketika input teks mencapai hingga 200 karakter, sistem menggunakan pendekatan penyaringan berbasis konteks. Gestur dipilih berdasarkan empat konteks percakapan utama seperti Salam, Perpisahan, Kesepakatan, dan Ketidaksepakatan. Metodologi ini terinspirasi dari karya **(Cassell dan Vilhjálmsson, 2019)**, yang menetapkan prinsip-prinsip dasar untuk memetakan fenomena percakapan ke perilaku komunikatif dalam agen *virtual*.



Gambar 2. Aturan Pada Dialog Pendek

Dalam Mode Dialog Panjang, ketika menangani input lebih dari 200 karakter, sistem beralih ke pendekatan penyaringan berbasis nada. Nilai 200 dalam string berasal dari perhitungan yang diambil dari karya **(Merdivan, 2020)**, yang mencatat dialog percakapan santai. Dalam penelitian ini, pada percakapan yang umum, manusia cenderung berbicara satu kalimat atau lebih, yang kira-kira setara dengan 200 karakter atau *string*. Oleh karena itu, nilai parameter untuk Mode Seleksi pada Dialog Panjang dan Pendek cukup mewakili. Gestur Utama dikategorikan dan dipilih berdasarkan tiga nada berbeda seperti Agresif, Netral, dan Damai. Kategorisasi ini dipengaruhi oleh penelitian **(Atmaja dan Sasou, 2022)** tentang pengenalan emosi dalam input suara, yang kemudian diadaptasi ke domain generasi gestur untuk dialog yang lebih panjang.



Gambar 3. Aturan Pada Dialog Panjang

2.3 Animasi Gestur dan *Motion Capture*

Aspek penting dalam menciptakan agen *virtual* yang meyakinkan dan menarik terletak pada kualitas serta kealamian gerakannya. Dalam proyek agen *virtual* Kemala, penulis sangat menekankan pada pencapaian animasi gestur yang realistis untuk meningkatkan interaksi dan imersi pengguna. Untuk mencapai hal ini, penulis memanfaatkan teknologi *motion capture* mutakhir, khususnya *Rokoko Vision*, yang terintegrasi dengan baik dengan *Unreal Engine* 5.

Rokoko Vision adalah alat *motion capture* canggih yang memungkinkan perekaman gerakan manusia dengan presisi yang cukup. Dengan memanfaatkan teknologi ini, penulis dapat menangkap gestur manusia yang cukup halus dengan fidelitas mumpuni, menerjemahkannya menjadi animasi yang mengalir dan tampak cukup hidup untuk agen *virtual* penulis.

Gambar 4. Contoh Gestur

Proses mengintegrasikan generasi gestur yang ditangkap dari gerakan manusia ke dalam Kemala melibatkan beberapa langkah berikut:

1. Perencanaan: Penulis mengidentifikasi dan mengkategorikan gestur kunci yang dibutuhkan untuk berbagai skenario interaksi, seperti sapaan, persetujuan, penolakan serta emosi-emosi spesifik seperti bahagia, netral dan agresif. Setiap gerakan dirancang agar sesuai dengan konteks dialog dan untuk mendukung makna verbal.
2. Pengambilan: Proses ini melibatkan seorang aktor yang melakukan gerakan gestur di dalam ruangan khusus bersama dengan perlengkapan Rokoko Vision. Untuk bekerja perangkat minimal Rokoko Vision adalah menggunakan 2 kamera untuk tampak depan dan samping aktor serta melakukan kalibrasi checkerboard untuk memastikan posisi aktor. Sistem ini mencatat setiap gerakan berdasarkan perekaman video pergerakan sang aktor dan diproses menjadi data playable yang bisa dilihat didalam software rokoko. Setiap sesi perekaman difokuskan pada satu jenis interaksi dan juga dialog, dengan tujuan untuk memastikan akurasi data untuk tiap skenario.
3. Pemrosesan: Data yang sudah diproses dan bisa dimainkan pada software rokoko kemudian dibersihkan dan dioptimalkan lebih lanjut menjadi file animasi sesuai kebutuhan agar bisa digunakan di software lainnya. Pada penelitian ini penulis menkonversinya menjadi format FBX *Unreal Engine* 5.
4. Integrasi: Animasi yang telah diproses diimpor ke dalam Unreal Engine 5 kemudian diterapkan pada model 3D Kemala. Dalam tahap ini, penyesuaian dilakukan untuk mengintegrasikan animasi dengan kerangka karakter Kemala secara optimal sehingga pergerakannya bisa semulus mungkin.
5. Pengujian: Penulis melakukan uji coba dalam berbagai konteks dialog yang akan dihadapi oleh Kemala. Hal ini meliputi pengaturan timing, kecepatan gerakan, dan penyesuaian transisi untuk memastikan bahwa setiap animasi sesuai. Penulis juga melanjutkan sinkronisasi gestur yang ada dengan model berbasis aturan yang sudah dibuat agar gestur sesuai dengan konteks emosional dan dialog. Pengujian ini memastikan agen virtual dapat merespons percakapan secara tepat, meningkatkan persepsi kealamian dan keterlibatan pengguna.

2.4 Studi Analisis Komparatif

Inti Penelitian penulis mengenai interaksi agen *virtual* melibatkan studi komparatif yang menyeluruh terhadap empat varian sistem generasi gestur yang berbeda. Pendekatan ini memungkinkan penulis untuk mengevaluasi efektivitas dan persepsi naturalitas dari sistem yang penulis kembangkan dibandingkan dengan berbagai kontrol. Keempat varian yang penulis teliti adalah:

1. Sistem Berbasis Aturan (RB): Ini adalah sistem utama yang penulis kembangkan, menggunakan serangkaian aturan yang sudah ditentukan untuk mengarahkan pemilihan gestur. Gestur dikelompokkan berdasarkan panjang dialog dan konteks percakapan tertentu, seperti salam dan perpisahan, serta nada emosional, seperti agresif, netral, dan damai. Pendekatan ini bertujuan menghasilkan gestur yang sesuai secara kontekstual dan terlihat alami, sehingga meningkatkan kelancaran dan realisme interaksi.
2. Gestur Ground Truth (GT): Varian ini berfungsi sebagai mekanisme kontrol, menggunakan data tangkapan gerakan manusia yang nyata. Sistem ini merupakan aplikasi langsung dari gerakan manusia yang direkam, menawarkan tingkat pergerakan yang paling alami dan memberikan dasar untuk membandingkan efektivitas sistem generasi gestur sintetis.
3. Gestur Acak (RG): Sistem ini memilih gestur secara acak dari *database* yang tersedia. Penyertaan sistem ini dalam studi memiliki beberapa tujuan. Pendekatan ini berfungsi

sebagai kontrol batas bawah untuk kesesuaian gestur, membantu mengevaluasi pentingnya konteks dalam pemilihan gestur.

Gambar 5. Perbedaan Gestur Akibat Perbedaan Model

2.5 Matriks Peniliaian

Penelitian penulis tentang sistem generasi gestur agen *virtual* Kemala menggunakan metodologi evaluasi yang terinspirasi kuat dari GENEA Challenge 2022, yang menjadi tolak ukur di bidang generasi gestur agen *virtual*. Evaluasi ini berfokus pada dua metrik utama yang penting untuk generasi gestur yang efektif pada agen *virtual*: *Human Likeness* dan *Appropriateness*. *Human Likeness* mengevaluasi seberapa mirip gestur yang dihasilkan dengan gerakan manusia alami, yang sangat penting untuk meningkatkan keterhubungan dan penerimaan terhadap agen *virtual*. Sementara itu, *Appropriateness* menilai seberapa baik gestur tersebut mendukung dan selaras dengan konten yang diucapkan, sehingga berpotensi meningkatkan komunikasi dan memberikan koherensi dalam interaksi.

Proses evaluasi melibatkan 43 peserta dari berbagai latar belakang demografis untuk meminimalkan bias. Peserta menonton video rekaman Kemala yang melakukan gestur dalam berbagai skenario, yang dirancang untuk mencakup berbagai konteks percakapan dan nada emosional. Video-video tersebut mencakup empat varian sistem generasi gestur yakni *Rule-Based* penulis, gestur *Ground Truth* (data gerakan manusia asli dari penangkapan gerak), *Random Gestures*, dan *Idling Gestures*. Pendekatan komprehensif ini memungkinkan analisis perbandingan sistem yang penulis kembangkan dengan berbagai kontrol. Para peserta memberikan penilaian terhadap gestur menggunakan skala Likert untuk *Human Likeness* dan *Appropriateness* segera setelah menonton setiap video. Skala ini biasanya berkisar dari 1 hingga 5, dengan deskripsi yang jelas untuk setiap poin. Misalnya, skala *Human Likeness* bisa berkisar dari 1 (Sangat Artifisial) hingga 10 (Sangat Mirip Manusia), sementara skala *Appropriateness* bisa dari 1 (Sama Sekali Tidak Tepat) hingga 10 (Sangat Tepat). Untuk mendapatkan wawasan tambahan, penulis juga menyertakan pertanyaan terbuka untuk umpan balik kualitatif.

Data yang dikumpulkan melalui proses ini dianalisis secara kuantitatif. Penulis melakukan analisis statistik pada penilaian skala Likert, termasuk pengukuran tendensi sentral dan variabilitas, serta analisis perbandingan di antara empat varian sistem. Proses perhitungan ini sesuai dengan proses perhitungan HEMVIP karya **(Jonell, dkk, 2021)**.

3. HASIL DAN PEMBAHASAN

Pada studi ini untuk memetakan hasil, penulis menggunakan *box plot*, *pie chart*, dan ANOVA untuk membandingkan rating rata-rata di berbagai sistem, dengan fokus pada kinerja sistem berbasis aturan penulis dalam skenario dialog pendek dan panjang.

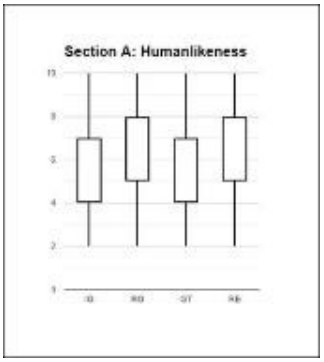
3.1 Evaluasi Sistem Berdasarkan *Human Likeness*

3.1.1. Dialog Pendek

Model Berbasis Aturan (RB) menunjukkan kinerja yang lebih baik dalam hal kemiripan manusia, mencapai *rating* median tertinggi sebesar 7. Skor yang lebih tinggi ini, bersama dengan sebaran rating yang lebih kecil, menunjukkan bahwa model RB secara konsisten menghasilkan gestur yang dianggap lebih mirip manusia oleh peserta. Distribusi rating yang sempit menunjukkan bahwa gestur dari model RB dinilai dengan variabilitas yang lebih sedikit, mencerminkan tingkat kepercayaan yang lebih tinggi dalam realisme manusia yang mereka tampilkan di antara para *evaluator* yang berbeda.

Menariknya, gestur *Ground Truth* (GT) tidak mendapatkan skor sebagus yang diharapkan, dengan rating median sebesar 5—mirip dengan Gestur *Idle* (IG). Hasil yang tidak terduga ini menunjukkan bahwa gestur alami, yang mungkin diperkirakan akan melampaui model sintetis, tidak selalu dianggap lebih unggul dalam hal kemiripan manusia. Kinerja GT dan IG yang sebanding menimbulkan pertanyaan tentang faktor-faktor yang memengaruhi persepsi manusia, seperti konteks gestur atau sifat interaksinya.

Model RB dan GT menunjukkan konsistensi yang lebih besar dalam rating mereka dibandingkan dengan Gestur *Idle* (IG) dan Gestur Acak (RG). Konsistensi ini menunjukkan bahwa pendekatan terstruktur, berbasis aturan (seperti model RB) dan gerakan manusia yang alami (yang diwakili oleh GT) cenderung menghasilkan hasil yang lebih dapat diandalkan dan dapat diulang dalam hal kemiripan manusia yang dipersepsikan. Sebaliknya, variabilitas yang lebih besar terlihat dalam rating untuk IG dan RG menunjukkan bahwa gestur acak atau non-sistematis memperkenalkan lebih banyak ketidakpastian dan inkonsistensi dalam persepsi kemiripan manusia.



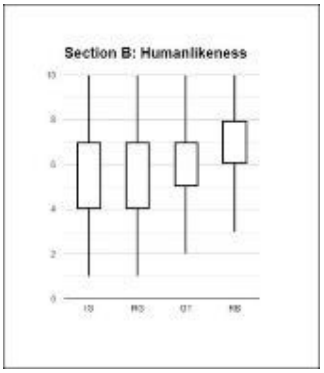
Gambar 6. Hasil Eksperimen *Humanlikeness* pada Dialog Pendek

Analisis kinerja model Berbasis Aturan (RB) mengungkapkan bahwa model ini secara konsisten menghasilkan gestur yang dianggap lebih mirip manusia, dengan rating median tertinggi sebesar 7 dan sebaran rating yang lebih kecil, menandakan efektivitasnya. Menariknya, gestur *Ground Truth* (GT), yang diharapkan dapat melampaui model sintetis, memiliki rating median sebesar 5, yang sebanding dengan Gestur *Idle* (IG), menantang asumsi bahwa gerakan manusia alami selalu lebih baik daripada gestur yang dihasilkan. Ini menunjukkan bahwa konteks atau jenis interaksi dapat berperan dalam persepsi manusia. Selain itu, baik model RB

maupun GT menunjukkan konsistensi yang lebih besar dalam rating dibandingkan dengan IG dan Gestur Acak (RG), menunjukkan bahwa pendekatan sistematis, baik berbasis aturan atau dihasilkan oleh manusia, menghasilkan hasil kemiripan manusia yang lebih dapat diandalkan. Sebaliknya, IG dan RG menunjukkan lebih banyak variabilitas dalam rating, menyoroti ketidakpastian dan persepsi alami yang berkurang dalam gestur non-sistematis. Secara keseluruhan, ini menunjukkan keunggulan model terstruktur seperti RB dalam menghasilkan gestur yang stabil dan mirip manusia.

3.1.2. Dialog Panjang

Dalam evaluasi pada dialog panjang, model Berbasis Aturan (RB) sekali lagi menunjukkan kinerja terbaik, dengan skor median 7 dan rata-rata tertinggi sebesar 7,0. Konsistensi ini menyoroti kemampuan model untuk menghasilkan gestur yang dianggap lebih mirip manusia dan sesuai untuk interaksi yang lebih lama. Stabilitas skor tinggi baik dalam median maupun rata-rata menunjukkan bahwa peserta menilai gestur yang dihasilkan oleh model RB tidak hanya lebih realistis, tetapi juga lebih cocok untuk kebutuhan emosional dan percakapan dalam dialog yang lebih panjang.



Gambar 7. Hasil Eksperimen *Humanlikeness* pada Dialog Panjang

Gestur Ground Truth (GT), meskipun tidak mendapatkan skor setinggi model RB, menunjukkan peningkatan yang signifikan dalam konteks dialog panjang. Model GT mencapai skor median 6, melampaui baik *Idle* Gestures (IG) maupun Random Gestures (RG). Peningkatan ini mengindikasikan bahwa gestur GT lebih disukai dalam konteks percakapan yang diperpanjang, di mana ekspresivitas dan kompleksitas yang mirip manusia menjadi lebih penting. Temuan ini mendukung gagasan bahwa gestur alami manusia cenderung lebih dihargai dalam lingkungan yang membutuhkan komunikasi non-verbal yang lebih berkelanjutan, meskipun masih kurang konsisten dibandingkan kinerja tinggi model RB.

Hasil statistik memberikan wawasan lebih lanjut tentang perbedaan signifikan antara jenis gestur. F-statistik sebesar 17,907 dan nilai p yang sangat rendah yaitu 8,27 x 10^-11 menunjukkan adanya perbedaan statistik yang kuat antara kinerja model RB, GT, IG, dan RG. Tingkat signifikansi statistik ini menyiratkan bahwa perbedaan yang diamati bukan disebabkan oleh variasi acak, tetapi mencerminkan perbedaan substansial dalam bagaimana peserta menilai model-model ini. Hasil ini menekankan dampak pendekatan berbasis aturan dalam generasi gestur, terutama dalam skenario dialog panjang yang lebih kompleks.

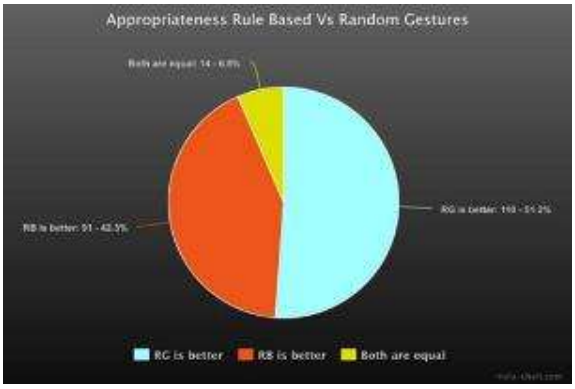
Model Berbasis Aturan terus unggul, dengan skor median dan rata-rata tertinggi, yang mengindikasikan kemampuannya menghasilkan gestur paling mirip manusia baik pada dialog pendek maupun panjang. Konsistensi penilaiannya menyoroti keuntungan dari pendekatan sistematis dalam generasi gestur, terutama dalam interaksi yang lebih panjang di mana

peserta mengharapkan tingkat nuansa emosional dan kesesuaian yang lebih tinggi. Meskipun gestur Ground Truth meningkat dalam dialog panjang, mereka masih kalah dibandingkan model RB, meskipun tetap lebih baik dari *Idle* dan Random Gestures. Peningkatan ini menunjukkan bahwa gestur manusia alami lebih dihargai dalam konteks yang lebih panjang di mana ekspresivitas menjadi kunci, namun mereka kurang konsisten dibandingkan metode berbasis aturan. Analisis statistik menegaskan bahwa perbedaan yang diamati antara jenis gestur sangat signifikan, yang semakin memperkuat efektivitas model berbasis aturan dalam menciptakan gestur yang sesuai dengan ekspektasi manusia dalam berbagai pengaturan dialog.

3.2 Evaluasi Sistem Berdasarkan *Appropriateness*

3.2.1. RB dan RG

Data menunjukkan bahwa 51,2 persen responden lebih menyukai *Random Gestures* (Gestur Acak), sementara 42,3 persen memilih gestur berbasis aturan (*Rule-Based*), dan 6,5 persen merasa keduanya sama-sama sesuai. Distribusi preferensi ini menunjukkan beberapa nuansa menarik dalam cara peserta memandang kesesuaian gestur. Meskipun desain sistem *Rule-Based* (RB) yang terstruktur dan konsisten, mayoritas responden lebih memilih sifat spontan dari *Random Gestures*. Ini bisa menunjukkan bahwa dalam konteks tertentu, peserta merasa bahwa ketidakpastian dan variasi dari *Random Gestures* lebih menarik atau alami daripada sistem berbasis aturan yang dirancang dengan hati-hati. Preferensi terhadap hal yang acak mungkin berasal dari kenyataan bahwa gestur manusia secara alami beragam, dan terkadang gerakan acak atau tidak terduga meniru variabilitas ini lebih baik daripada sistem berbasis aturan.



Gambar 8. Hasil Eksperimen *Appropriateness* antara RB dan RG

Adapun 6,5 persen yang melihat *Random Gestures* dan *Rule-Based Gestures* sama-sama sesuai mungkin menunjukkan bahwa bagi beberapa orang, perbedaan antara kedua pendekatan ini tidak cukup signifikan untuk menimbulkan preferensi yang kuat. Kelompok ini mungkin melihat kedua gaya gestur cocok dalam konteks tertentu, yang menyiratkan bahwa dalam kondisi tertentu, kelebihan dari gestur berbasis aturan yang sistematis dan spontanitas dari gestur acak bisa saling melengkapi.

Mayoritas responden, yaitu 51,2 persen, lebih menyukai *Random Gestures* dibandingkan gestur berbasis aturan, menunjukkan bahwa ketidakpastian lebih mencerminkan variasi alami yang terlihat dalam perilaku manusia. Preferensi ini mencerminkan gagasan bahwa dalam beberapa konteks, hal yang acak mungkin terasa lebih spontan dan hidup, karena gestur manusia jarang sepenuhnya terstruktur. Namun, 42,3 persen yang memilih gestur berbasis aturan menandakan bahwa banyak peserta tetap menghargai konsistensi dan kesesuaian kontekstual dari model yang sistematis. Sementara itu, 6,5 persen yang menganggap kedua

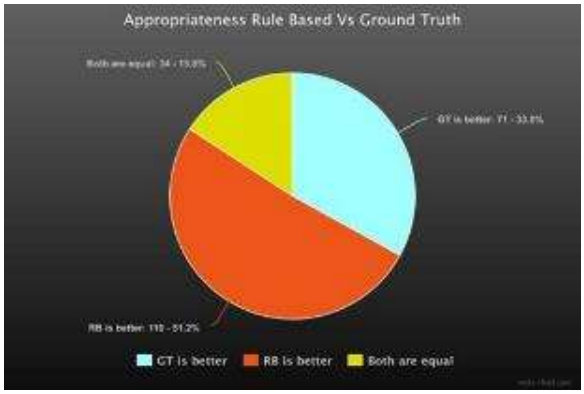
pendekatan ini sama-sama sesuai menunjukkan bahwa dalam situasi tertentu, perbedaan antara *Random Gestures* dan *Rule-Based Gestures* menjadi kabur, mengindikasikan bahwa kedua pendekatan tersebut bisa dianggap efektif tergantung pada konteks percakapan.

Temuan ini menyoroti pentingnya menyeimbangkan spontanitas alami dengan generasi gestur yang terstruktur untuk pengalaman pengguna yang optimal. Namun, 42,3 persen peserta yang lebih menyukai gestur berbasis aturan mencerminkan bahwa sebagian besar audiens masih menghargai konsistensi, presisi, dan kesesuaian konteks—ciri khas pendekatan berbasis aturan. Peserta ini mungkin merasa gestur berbasis aturan lebih selaras dengan percakapan, memperkuat gagasan bahwa model terstruktur dapat meniru perilaku komunikasi manusia dengan baik.Analisis kinerja model Berbasis Aturan (RB) mengungkapkan bahwa model ini secara konsisten menghasilkan gestur yang dianggap lebih mirip manusia, dengan rating median tertinggi sebesar 7 dan sebaran rating yang lebih kecil, menandakan efektivitasnya.

Menariknya, gestur Ground Truth (GT), yang diharapkan dapat melampaui model sintetis, memiliki rating median sebesar 5, yang sebanding dengan Gestur *Idle* (IG), menantang asumsi bahwa gerakan manusia alami selalu lebih baik daripada gestur yang dihasilkan. Ini menunjukkan bahwa konteks atau jenis interaksi dapat berperan dalam persepsi manusia. Selain itu, baik model RB maupun GT menunjukkan konsistensi yang lebih besar dalam rating dibandingkan dengan IG dan Gestur Acak (RG), menunjukkan bahwa pendekatan sistematis, baik berbasis aturan atau dihasilkan oleh manusia, menghasilkan hasil kemiripan manusia yang lebih dapat diandalkan. Sebaliknya, IG dan RG menunjukkan lebih banyak variabilitas dalam rating, menyoroti ketidakpastian dan persepsi alami yang berkurang dalam gestur non-sistematis. Secara keseluruhan, ini menunjukkan keunggulan model terstruktur seperti RB dalam menghasilkan gestur yang stabil dan mirip manusia.

3.2.2. RB dan GT

Data menunjukkan bahwa 51.2 persen responden lebih memilih gestur Berbasis Aturan (*Rule-Based*), sementara 33.0 persen memilih gestur *Ground Truth*, dan 15.8 persen menganggap keduanya sama-sama sesuai. Distribusi preferensi ini mencerminkan kekuatan model *Rule-Based* (RB) dalam menghasilkan gestur yang dianggap peserta lebih menyerupai manusia atau lebih sesuai dengan konteks. Dengan lebih dari setengah responden memilih gestur RB, ini menunjukkan bahwa pendekatan yang terstruktur dan diatur oleh aturan dalam generasi gestur menawarkan prediktabilitas dan kesesuaian yang terasa pas di mata pengguna. Kemungkinan besar karena gestur RB dirancang dengan cermat agar sesuai dengan konteks dialog dan isyarat emosional tertentu, sehingga interaksi yang dihasilkan terasa lebih kohesif dan sesuai.



Gambar 9. Hasil Eksperimen *Appropriateness* antara RB dan GT

Sebanyak 15.8 persen yang menganggap keduanya sama-sama sesuai mengindikasikan bahwa bagi sebagian peserta, perbedaan antara gestur berbasis aturan dan gestur manusia nyata mungkin tidak cukup menonjol untuk memunculkan preferensi yang kuat. Hal ini menunjukkan bahwa dalam beberapa kasus, pendekatan terstruktur dari gestur berbasis aturan dapat meniru spontanitas dan aliran alami gerakan manusia, sehingga membuat kedua sistem tersebut dianggap sama-sama valid tergantung pada konteksnya.

Mayoritas responden, yaitu 51.2 persen, memilih gestur Berbasis Aturan, menyoroti kemampuan sistem untuk menghasilkan gerakan yang lebih sesuai dengan konteks dan lebih menyerupai gerakan manusia. Preferensi ini menunjukkan bahwa pendekatan terstruktur menawarkan konsistensi dan relevansi yang sejalan dengan ekspektasi peserta dalam interaksi gestur. Sementara itu, 33.0 persen memilih gestur *Ground Truth*, mencerminkan keinginan untuk kualitas yang lebih alami dan kurang terstruktur yang ditawarkan oleh gestur manusia. Sebanyak 15.8 persen yang menganggap keduanya sama-sama sesuai menunjukkan bahwa bagi beberapa pengguna, perbedaan antara ketepatan terstruktur dari gestur berbasis aturan dan keaslian gerakan manusia tidak terlalu jelas, menunjukkan potensi sistem berbasis aturan untuk meniru spontanitas yang menyerupai gerakan manusia dengan efektif. Fakta bahwa 33.0 persen lebih menyukai gestur *Ground Truth* menunjukkan bahwa sebagian besar responden masih menghargai keaslian dan kualitas alami yang ditemukan pada gestur manusia nyata. Meskipun gestur *Ground Truth* mungkin kurang konsisten dalam kinerjanya dibandingkan sistem berbasis aturan, mereka membawa kesan mentah dan organik yang menarik bagi pengguna yang mencari gestur yang lebih selaras dengan gerakan alami manusia.

3.3 Dampak Panjang Dialog

Perbedaan kinerja yang diamati antara *Ground Truth* (GT) dan gestur *Rule-Based* (RB) dalam dialog pendek dan panjang mengungkapkan wawasan penting tentang persepsi dan ekspektasi manusia terhadap komunikasi non-verbal pada agen *virtual*. Gestur GT, yang berasal dari gerakan manusia sebenarnya, tampil lebih baik dalam dialog panjang dibandingkan dialog pendek. Hal ini sejalan dengan penelitian pola komunikasi manusia. Misalnya, **(Kendon, dkk, 1980)** mengamati bahwa manusia cenderung menggunakan lebih banyak gestur dan urutan gestur yang lebih kompleks dalam tindakan bicara yang diperpanjang. Demikian pula, **(Arnheim, dkk, 1994)** mencatat bahwa tingkat dan kompleksitas gestur sering meningkat seiring dengan durasi dan kompleksitas pembicaraan. Ini menunjukkan bahwa partisipan dalam studi kami mungkin secara tidak sadar mengharapkan gestur yang lebih hidup dan bervariasi selama dialog yang lebih panjang, mencerminkan perilaku manusia alami.

Peningkatan persepsi terhadap GT dalam dialog panjang juga sesuai dengan konsep "*gesture units*" yang diperkenalkan oleh **(Kendon, dkk, 1980)**. Dia mengusulkan bahwa gestur sering muncul dalam frasa atau unit yang selaras dengan struktur pembicaraan. Dalam dialog yang lebih panjang, terdapat lebih banyak kesempatan bagi unit gestur ini untuk berkembang sepenuhnya dan dipersepsikan, yang pada akhirnya menghasilkan tampilan yang lebih alami. Di sisi lain, sistem *Rule-Based* (RB) kami menunjukkan kinerja yang konsisten di kedua dialog pendek maupun panjang. Adaptabilitas ini menunjukkan bahwa aturan yang mengatur sistem kami berhasil menangkap aspek-aspek fundamental dari gerakan gestur manusia yang berlaku terlepas dari panjang dialog. Konsistensi kinerja RB ini sejalan dengan temuan **(Cassell, dkk, 1999)**, yang menunjukkan bahwa sistem berbasis aturan dapat menghasilkan gestur yang sesuai dalam berbagai konteks diskusi.

Adaptabilitas sistem RB kami mungkin disebabkan oleh desainnya, yang kemungkinan menggabungkan prinsip-prinsip dari komunikasi manusia dalam bentuk pendek maupun

panjang. Pendekatan ini didukung oleh **(Kopp, dkk, 2006)**, yang menekankan pentingnya mempertimbangkan faktor lokal (jangka pendek) dan global (jangka panjang) dalam sistem generasi gestur.

Lebih lanjut, kinerja konsisten dari RB pada berbagai panjang dialog menunjukkan bahwa sistem ini berhasil menyeimbangkan antara gestur minimal yang diharapkan dalam ujaran pendek dan gestur yang lebih rumit dalam ujaran yang lebih panjang. Keseimbangan ini sangat penting dalam menciptakan agen *virtual* yang meyakinkan, seperti yang dicatat oleh **(Kipp 2004)** dalam karyanya tentang sintesis gestur.

3.4 Diskusi

Sistem RB (*rule-based*) yang kami kembangkan menunjukkan performa yang luar biasa dalam menciptakan gestur yang menyerupai manusia, baik dalam dialog pendek maupun panjang, bahkan sering kali mengungguli gestur GT (*ground truth*), yang merupakan gestur sebenarnya dari data manusia. Hal ini menegaskan bahwa pendekatan kami mampu menangkap aspek-aspek penting dari gestur manusia secara efektif. Salah satu temuan utama dari penelitian ini adalah adanya pengaruh signifikan dari panjang dialog terhadap persepsi gestur. Partisipan dalam penelitian ini cenderung mengharapkan lebih banyak gestur pada dialog yang lebih panjang, dan sistem RB kami mampu menyesuaikan diri dengan ekspektasi ini. Ini menunjukkan bahwa sistem yang dirancang dengan pendekatan berbasis aturan (*rule-based*) dapat menangani variasi dalam dialog panjang dengan baik, memberikan gestur yang lebih banyak dan kompleks yang sesuai dengan dinamika percakapan tersebut.

Selain itu, meskipun sistem RB mampu menghasilkan gestur yang sangat mirip dengan manusia, dalam dialog pendek, ada sedikit kecenderungan partisipan untuk lebih memilih gestur RG (*random-generated*) dalam hal kesesuaian konteks. Hal ini menunjukkan bahwa dalam generasi gestur, ada kebutuhan untuk menyeimbangkan antara prediktabilitas yang dihasilkan oleh mekanisme berbasis aturan dengan variasi yang lebih dinamis yang sering ditemukan dalam gestur acak. Prediktabilitas yang terlalu tinggi dalam sistem berbasis aturan dapat membuat gestur tampak terlalu terencana atau kaku, terutama dalam interaksi pendek yang mungkin memerlukan lebih banyak spontanitas. Oleh karena itu, pendekatan ini menyarankan bahwa, meskipun gestur RB berhasil memberikan ilusi gerakan yang mirip dengan manusia, untuk dialog-dialog pendek yang membutuhkan fleksibilitas, sistem berbasis aturan masih perlu dikombinasikan dengan elemen-elemen yang lebih acak agar hasilnya terlihat lebih natural.

Menariknya, penelitian ini juga menemukan bahwa sistem RB kami secara konsisten mengungguli gestur GT, khususnya dalam konteks dialog pendek. Temuan ini mungkin terdengar mengejutkan, namun menyoroti fakta bahwa gestur yang dirancang secara sistematis melalui aturan dapat, dalam beberapa kasus, lebih efektif dalam menciptakan ilusi gerakan manusia daripada gestur manusia asli ketika diterapkan di lingkungan *virtual*. Hal ini mungkin terjadi karena sistem RB mampu secara presisi menyesuaikan gerakan dengan konteks percakapan yang ada, sementara gestur manusia asli kadang kala kurang terstruktur atau tidak sepenuhnya selaras dengan kebutuhan interaksi *virtual*. Dalam ruang *virtual*, di mana segala sesuatunya sering kali lebih terkendali dan membutuhkan kesesuaian yang lebih besar dengan stimulus visual dan verbal, gestur yang dihasilkan secara sistematis mungkin lebih terlihat "manusiawi" dibandingkan gestur alami yang kurang terkoordinasi.

Dari hasil-hasil ini, tampak jelas bahwa ada potensi besar untuk mengembangkan sistem hibrida yang menggabungkan kekuatan prediktabilitas dari mekanisme berbasis aturan dengan fleksibilitas dan spontanitas dari gestur acak. Sistem hibrida semacam ini akan memungkinkan generasi gestur yang lebih bervariasi dan dinamis, menyesuaikan diri dengan kompleksitas

serta keragaman interaksi manusia. Sistem hibrida yang menggabungkan *data driven model dan rule-based,* sebelumnya juga pernah diteliti pada karya milik **(Sadoughi, 2018)** yang terbukti memang menghasilkan gestur yang lebih sesuai. Walau begitu sistem ini tetap memiliki kekurangan dalam percakapan jangka panjang. Menurut permasalahan terkait pengembangan berkelanjutan menjadi poin penting untuk beralih dari sistem hibrida dan menggunakan metode regresi yang cenderung lebih flexibel.

4. KESIMPULAN

Penelitian ini mengevaluasi efektivitas berbagai model generasi gestur, dengan fokus khusus pada sistem berbasis aturan (Rule-Based/RB) dibandingkan dengan Ground Truth (GT), Random Gestures (RG), dan Idling Gestures (IG). Temuan kami menunjukkan bahwa sistem berbasis aturan secara konsisten menghasilkan gestur yang dianggap lebih menyerupai manusia baik dalam dialog pendek maupun panjang. Menariknya, meskipun gestur GT diharapkan mengungguli model sintetis karena asalnya yang alami, kinerjanya tidak sebaik yang diperkirakan dalam dialog pendek. Selain itu, preferensi partisipan terhadap spontanitas RG dalam situasi tertentu mengindikasikan perlunya keseimbangan antara struktur dan variabilitas alami, serta potensi integrasi model berbasis aturan dengan pendekatan yang lebih fleksibel untuk meningkatkan realisme agen *virtual*.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Universitas Indonesia yang mendukung penelitian ini melalui Hibah Publikasi Internasional Terindeks (PUTI) Q2, 2023, nomor: NKB-804/UN2.RST/HKP.05.00/2023.

DAFTAR RUJUKAN

Agarwal, A. (2023, April 12). Unreal Engine and its Evolution | Extern Labs Inc. Extern Labs Blog | Delivering IT Innovation. Extern Labs.

Arnheim, R., & McNeill, D. (1994). Hand and Mind: What Gestures Reveal about Thought. Leonardo, *27*(4), 358. https://doi.org/10.2307/1576015

Atmaja, B. T., & Sasou, A. (2022). Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations. *Sensors*, *22*(17), 6369. https://doi.org/10.3390/s22176369

Calvaresi, D., Eggenschwiler, S., Mualla, Y., Schumacher, M., & Calbimonte, J.-P. (2023). Exploring agent-based chatbots: a systematic literature review. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 11207–11226. https://doi.org/10.1007/s12652-023-04626-5

Cassell, J. (2001). Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine*, *22*(4), 67. https://doi.org/10.1609/aimag.v22i4.1593

Cassell, J., & Vilhjálmsson, H. (1999). Fully Embodied Conver- sational Avatars: Making Communicative Behaviors Autonomous. Autonomous Agents and Multi-agent Systems,. *Autonomous Agents and Multi-Agent Systems*, *2*(1), 45–64. https://doi.org/10.1023/A:1010027123541

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*, (pp. 93–98). https://doi.org/10.1145/3267851.3267898

Ferstl, Y., Neff, M., & McDonnell, R. (2019). Multi-objective adversarial gesture generation. Motion, *Interaction and Games*, (pp. 1–10). https://doi.org/10.1145/3359566.3360053

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., & Malik, J. (2019). Learning Individual Styles of Conversational Gesture. CoRR, abs/1906.04160. http://arxiv.org/abs/1906.04160

Gu, X., Yu, T., Huang, J., Wang, F., Zheng, X., Sun, M., Ye, Z., & Li, Q. (2023). Virtual-Agent-Based Language Learning: A Scoping Review of Journal Publications from 2012 to 2022. Sustainability, *15*(18), 13479. https://doi.org/10.3390/su151813479

HOSTETTER, A. B., & ALIBALI, M. W. (2008). Visible embodiment: Gestures as simulated action. Psychonomic *Bulletin & Review*, *15*(3), 495–514. https://doi.org/10.3758/PBR.15.3.495

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*, (pp. 707–711). https://doi.org/10.1145/3462244.3479957

Kendon, A. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. *In The Relationship of Verbal and Nonverbal Communication,* (pp. 207–228). DE GRUYTER MOUTON. https://doi.org/10.1515/9783110813098.207

Kim, Y., & Baylor, A. L. (2016). Research-Based Design of Pedagogical Agent Roles: a Review, Progress, and Recommendations. *International Journal of Artificial Intelligence in Education*, *26*(1), 160–169. https://doi.org/10.1007/s40593-015-0055-y

Kipp, M. (2004). Gesture Generation by Imitation: From Human Behavior to Computer Character Animation.

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*, (pp. 205–217). https://doi.org/10.1007/11821830_17

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds, 15(1), 39–52. https://doi.org/10.1002/cav.6

Krämer, N. C., Rosenthal-von der Pütten, A. M., & Hoffmann, L. (2015). Social Effects of Virtual and Robot Companions. *In The Handbook of the Psychology of Communication Technology,* (pp. 137–159). Wiley. https://doi.org/10.1002/9781118426456.ch6

Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., & Kjellström, H. (2021). Moving Fast and Slow: Analysis of Representations and Post-Processing in Speech-Driven Automatic Gesture Generation. International Journal of Human–Computer Interaction, *37*(14), 1300–1316. https://doi.org/10.1080/10447318.2021.1883883

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. *ACM Transactions on Graphics*, *43*(3), 1–28. https://doi.org/10.1145/3656374

Martin, A. (2024). ElevenLabs Review 2024 — Pricing, Features, and Alternatives. Technopedia. https://www.techopedia.com/ai/elevenlabs-review

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., & Geist, M. (2020). Human Annotated Dialogues Dataset for Natural Conversational Agents. *Applied Sciences*, *10*(3), 762. https://doi.org/10.3390/app10030762

Sadoughi, N., & Busso, C. (2018). Novel Realizations of Speech-Driven Head Movements with Generative Adversarial Networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), (pp. 6169–6173). https://doi.org/10.1109/ICASSP.2018.8461967

Tipper, C. M., Signorini, G., & Grafton, S. T. (2015). Body language in the brain: constructing meaning from expressive movement. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00450

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. CoRR, abs/1706.03762. http://arxiv.org/abs/1706.03762