Síntese Gestual Zero-Shot: Transferência de Estilo Comportamental Multimodal de Poses Tridimensionais

Zero-Shot Gesture Synthesis: Multimodal Behavioral Style Transfer from 3D Poses

Anna Carolina Souza Bispo[1,2], Breno Ramon Santana dos Santos[1], Erick Marck de Barros Menezes[1,2,3], Lívia Pereira Guimarães[1], Melyssa Maria Macedo Tatum[1], Rafael José Mecenas Silva[1,2], Victor Flávio Araujo[1,2,3]

[1]Grupo de Pesquisa Interdisciplinar em Tecnologia, Computação e Sociedade (GPITCS)
[2] Universidade Tiradentes (UNIT), Aracaju, SE – Brasil
[3] National Institute of Science and Technology Social and Affective Neuroscience (INCT-SANI)

{anna.bispo, breno.rsantana, erick.marck, livia.pguimaraes,
melyssa.maria, rafael.mecenas, victor.flavio93}@souunit.com.br

Abstract. Introduction*: This work presents a multimodal synthesis model for generating three-dimensional body gestures from speech and text data.* ***Objective:*** *The objective of this work is to design and evaluate a multimodal gesture synthesis model capable of generating 3D movements from speech and text, enabling zero-shot style transfer between different speakers, aiming to decouple semantic content from speaker style.* ***Methodology or Steps****: The neural architecture, based on GANs, employs a content encoder, a style encoder, and a sequence-to-sequence generator to effectively decouple semantic content from speaker stylistic features. The model processes acoustic features from Mel-spectrograms and semantic representations from BERT vectors, utilizing a physically grounded reconstruction loss with a differentiable Forward Kinematics (FK) layer to minimize joint positional error in 3D space, combined with a continuous 6D rotation representation for stable movements.* ***Results:*** *Experiments demonstrate the generation of coherent gestures, validated through quantitative metrics and qualitative analysis, confirming the model's ability to transfer style while preserving semantic intent.*
Keywords *Gesture synthesis, Gesture generation, Generative Adversarial Networks (GANs), 3D movements, Style transfer.*

Resumo. Introdução: *Este trabalho apresenta um modelo de síntese multimodal para a geração de gestos corporais tridimensionais a partir de dados de fala e texto.* ***Objetivo:*** *O objetivo deste trabalho é projetar e avaliar um modelo de síntese gestual multimodal capaz de gerar movimentos 3D a partir de fala e texto, permitindo a transferência de estilo zero-shot entre diferentes falantes, visando desacoplar o conteúdo semântico do estilo do locutor.* ***Metodologia ou Etapas****: A arquitetura neural, baseada em GANs, emprega um codificador de conteúdo, um codificador de estilo e um gerador sequência-a-sequência para desacoplar efetivamente o conteúdo semântico das características estilísticas do falante. O modelo processa características acústicas de Mel-espectrogramas e representações semânticas de vetores*

BERT, utilizando uma perda de reconstrução fisicamente fundamentada com uma camada diferenciável de Forward Kinematics (FK) para minimizar o erro posicional das articulações no espaço 3D, combinada com uma representação de rotação 6D contínua para movimentos estáveis. ***Resultados:*** *Os experimentos demonstram a geração de gestos coerentes, validados por métricas quantitativas e análise qualitativa, confirmando a capacidade do modelo de transferir estilo preservando a intenção semântica.*
Palavras-Chave *Síntese de gestos, Geração de gestos, Redes Adversariais Generativas (GANs), Movimentos 3D, Transferência de estilo.*

1. Introdução

A criação de personagens virtuais críveis e expressivos é um pilar para a imersão em jogos digitais e experiências interativas. A comunicação humana, no entanto, é intrinsecamente multimodal, onde gestos corporais não apenas acompanham a fala, mas também transmitem nuances, emoções e traços de personalidade [Mcneill 1994]. Gerar movimentos expressivos [Campbell-Kibler 2009, Obin et al. 2011] que sejam sincronizados com a prosódia da fala e que reflitam o estilo individual de um falante [Bishop 2006] representa um desafio significativo [Kucherenko et al. 2021]. A principal dificuldade reside em desacoplar o *o quê* está sendo dito (conteúdo semântico) do *como* está sendo dito (estilo comportamental).

Para endereçar essa lacuna, este artigo propõe um modelo de síntese gestual multimodal que utiliza Redes Adversariais Generativas (GANs) para gerar movimentos 3D a partir de dados de fala e texto. O objetivo central é alcançar a transferência de estilo *zero-shot*, permitindo que o estilo gestual de um locutor de referência seja aplicado a um conteúdo de fala arbitrário, produzindo animações expressivas e personalizadas. As principais contribuições deste trabalho são: (1) uma função de perda fisicamente fundamentada que utiliza uma camada diferenciável de Cinemática Direta (FK) para otimizar as posições 3D das articulações, garantindo coerência cinemática; (2) um pipeline de pré-processamento que emprega representações de rotação 6D para movimentos estáveis; e (3) o desacoplamento efetivo de conteúdo e estilo através de uma arquitetura adversária, validado no dataset *Trinity Speech-Gesture II* [Ferstl et al. 2021].

2. Trabalhos Relacionados

A síntese automática de gestos é uma área multidisciplinar que integra visão computacional, processamento de linguagem natural e animação digital [Kebe et al. 2024]. Sua evolução partiu de sistemas baseados em regras, como o pioneiro BEAT [Cassell et al. 2001], para abordagens modernas orientadas por dados. Revisões metodológicas ressaltam a importância da congruência entre modalidades para a construção de agentes interativos naturalistas [Lugrin et al. 2022]. Trabalhos iniciais de aprendizado profundo utilizaram modelos *seq2seq* com LSTMs para mapear características acústicas a sequências de movimento [Yoon et al. 2018]. A subsequente fusão com dados textuais demonstrou enriquecer a semântica dos gestos gerados, resultando em movimentos mais apropriados ao conteúdo verbal [Ahuja et al. 2020]. Autores como Kucherenko et al. [Kucherenko et al. 2021] forneceram um panorama abrangente das técnicas, destacando os desafios em avaliação e na geração de movimentos naturalistas.

No entanto, um desafio central que persiste é o desacoplamento entre o conteúdo da fala e o estilo do falante. Modelos recentes como o proposto por Ferstl et al. [Ferstl et al. 2021] e Mehta et al. [Mehta et al. 2024] avançaram ao propor arquiteturas unificadas para a síntese de fala e gestos. Contudo, a transferência de estilo explícita e *zero-shot* — aplicar um estilo aprendido a um conteúdo nunca visto — continua sendo uma área de pesquisa ativa. Nosso trabalho se insere neste contexto, buscando diferenciar-se pelo uso de uma arquitetura adversária projetada especificamente para isolar um vetor de estilo multimodal e pela aplicação de uma perda de reconstrução no espaço 3D, garantindo maior plausibilidade física aos movimentos, uma abordagem relevante para aplicações em jogos e agentes virtuais

3. Metodologia

A metodologia abrange o pré-processamento de dados multimodais, a arquitetura do modelo e a estratégia de treinamento. Os experimentos foram realizados no Google Colaboratory, utilizando PyTorch e o dataset público *Trinity Speech-Gesture II* [Ferstl et al. 2021], que contém áudio, texto e captura de movimento 3D sincronizados.

Pré-processamento e Representação de Dados. Para garantir estabilidade e coerência, desenvolvemos um pipeline de pré-processamento robusto. Para o **movimento**, evitamos o *gimbal lock* convertendo as rotações de cada uma das 69 articulações para a representação 6D contínua [Zhou et al. 2019], concatenada com o vetor de velocidade da articulação raiz. Isso resulta em um vetor de pose de dimensão 418 por quadro. Para o **áudio**, segmentos de 16 segundos são convertidos em espectrogramas Mel. Para o **texto**, as transcrições são processadas pelo modelo BERT para extrair um vetor de embedding semântico, calculado como a média ponderada dos estados ocultos da última camada, conforme a Equação 1.

$$E_{\text{text}} = \frac{\sum_{i=1}^{L} h_i \cdot M_i}{\sum_{i=1}^{L} M_i} \tag{1}$$

Onde h_i é o estado oculto, M_i a máscara de atenção e L o comprimento da sequência.

Arquitetura do Modelo. A arquitetura GAN proposta é composta por um Gerador e um Discriminador de Estilo. O **Gerador** possui três componentes: um *Codificador de Conteúdo*, que funde as características de áudio e texto usando CNNs e uma LSTM bidirecional para extrair a semântica; um *Codificador de Estilo*, que extrai um vetor de estilo de 256 dimensões a partir dos dados de entrada; e um *Gerador de Gestos*, um módulo *seq2seq* com LSTMs que recebe o conteúdo latente e o vetor de estilo para produzir a sequência final de poses. O **Discriminador de Estilo** é treinado para classificar o ID do falante a partir do conteúdo latente gerado. O treinamento adversário força o Gerador a remover informações de estilo do conteúdo, isolando-as no vetor de estilo.

Treinamento e Função de Perda. O modelo foi treinado por 100 épocas com o otimizador Adam. A função de perda do gerador, L_G, combina uma perda de reconstrução (L_{recon}) com uma perda adversarial (L_{adv}), ponderada por $\lambda_{\text{adv}} = 1.0$.

$$L_G = L_{\text{recon}} + \lambda_{\text{adv}} \cdot L_{\text{adv}} \tag{2}$$

A inovação central está na L_{recon}, que é fisicamente fundamentada. As rotações 6D geradas são passadas por uma camada de **Cinemática Direta (FK) diferenciável**, que

as converte em posições 3D globais das articulações. A perda é então calculada como a soma da perda L_1 entre as posições 3D geradas e as de referência (L_{pos}), e a perda L_1 das velocidades da raiz (L_{vel}).

$$L_{\text{recon}} = L_{\text{pos}} + 1.0 \cdot L_{\text{vel}} \tag{3}$$

Esta abordagem força o modelo a aprender a gerar movimentos que não apenas se parecem corretos, mas que são cinematicamente válidos no espaço 3D.

4. Resultados e Discussão

A avaliação do modelo combina análises quantitativas e qualitativas para verificar a fidelidade da reconstrução e a eficácia da transferência de estilo.

A análise quantitativa, baseada no Erro Quadrático Médio (MSE) entre as poses 6D geradas e as de referência, resultou em um **MSE final de 0.017326** após 100 épocas. Embora essa métrica agregada não capture nuances estilísticas, um valor baixo indica uma reconstrução geral precisa da estrutura do movimento. A evolução das perdas (Figura 1) corrobora a estabilidade do treinamento: a perda de reconstrução ($recon\_loss$) converge de forma consistente, indicando que o modelo aprendeu a gerar poses fisicamente plausíveis, enquanto a dinâmica competitiva entre as perdas do gerador e do discriminador adversarial (G_adv_loss e D_sty_loss) evidencia que o desacoplamento de estilo estava ocorrendo como esperado.

Na análise qualitativa, a Figura 2 compara uma pose de referência do dataset (a) com uma pose gerada pelo modelo (b). Observa-se que a macroestrutura do gesto — a elevação dos braços e a postura geral do tronco — é preservada, mantendo a intenção comunicativa do movimento original. Contudo, é visível uma leve suavização nos detalhes finos, como a articulação das mãos, um artefato comum em modelos generativos que otimizam perdas baseadas em média, como L1/MSE. O resultado mais significativo, no entanto, é a capacidade de transferência de estilo. Ao fornecer o mesmo conteúdo de fala (áudio e texto) mas condicionar a geração com vetores de estilo de diferentes falantes, o modelo produz sequências gestuais visivelmente distintas, que herdam a dinâmica e a amplitude de movimento do falante de estilo. Isso valida a eficácia da arquitetura adversária em isolar as características estilísticas, cumprindo o objetivo central do trabalho.

5. Considerações Finais

Este trabalho apresentou um modelo de síntese gestual capaz de realizar transferência de estilo *zero-shot* em movimentos 3D multimodais. A arquitetura adversária, combinada com função de perda baseada em Cinemática Direta, mostrou-se eficaz para desacoplar conteúdo e estilo. O modelo gera animações coerentes que preservam a intenção da fala, adotando características estilísticas de um locutor de referência. É contribuição relevante para criar personagens não-jogáveis (NPCs) e avatares naturais e personalizáveis em jogos e realidade virtual, onde o estilo é crucial para expressividade e credibilidade.

Como trabalhos futuros, planeja-se explorar perdas perceptuais para capturar detalhes de alta frequência, como gestos de mão finos, e incorporar atenção para melhorar a sincronização entre picos prosódicos da fala e gestos.

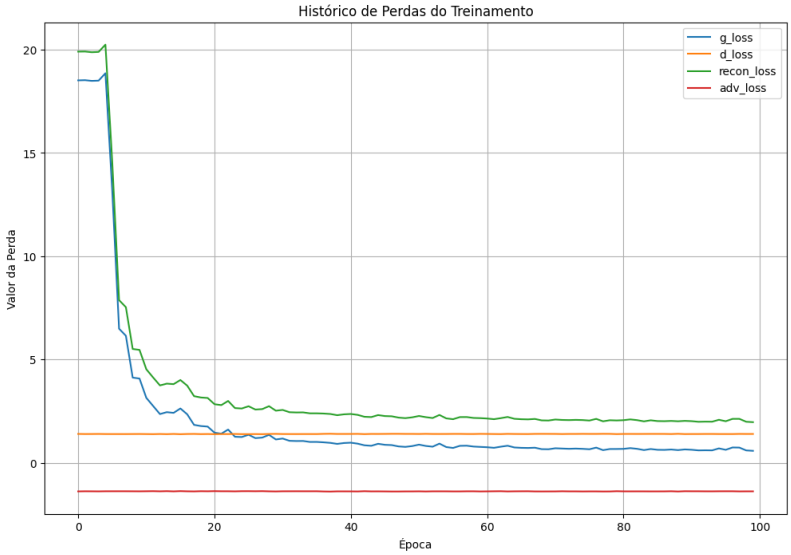
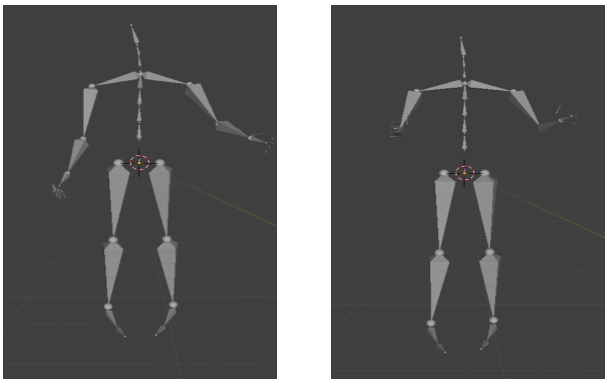


Figura 1. Curvas de perda durante o treinamento. A queda consistente da perda de reconstrução ($recon\_loss$) indica o aprendizado da geração de poses, enquanto as perdas adversariais (G_adv_loss, D_sty_loss) mostram a dinâmica de desacoplamento de estilo.



(a) Pose de referência do dataset

(b) Pose correspondente gerada

Figura 2. Análise qualitativa visual: (a) pose de referência extraída do dataset. (b) pose gerada pelo modelo para o mesmo instante de tempo, demonstrando a preservação da intenção do movimento.

Referências

Ahuja, C., Lee, D. W., Ishii, R., e Morency, L.-P. (2020). No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1884–1895.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*, volume 16. Springer.

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change*, 21(1):135–156.

Cassell, J., Vilhjálmsson, H., e Bickmore, T. (2001). Beat: the behavior expression animation toolkit. *ACM SIGGRAPH*, 2001:477–486.

Ferstl, Y., Neff, M., e McDonnell, R. (2021). Expressgesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, page e2016.

Kebe, G. Y., Birlikci, M. D., Boudin, A., Ishii, R., Girard, J. M., e Morency, L.-P. (2024). Gestics: A multimodal corpus for studying gesture synthesis in two-party interactions with contextualized speech. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '24)*, page 10, Glasgow, United Kingdom. ACM.

Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G. E., e and, H. K. (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.

Lugrin, B., Pelachaud, C., e Traum, D. (2022). *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*. ACM.

Mcneill, D. (1994). Hand and mind: What gestures reveal about thought. *Bibliovault OAI Repository, the University of Chicago Press*, 27.

Mehta, S., Tu, R., Alexanderson, S., Beskow, J., Székely, , e Henter, G. E. (2024). Unified speech and gesture synthesis using flow matching. *arXiv preprint arXiv:2310.05181*.

Obin, N., Lacheret, A., e Rodet, X. (2011). Stylization and trajectory modelling of short and long term speech prosody variations. In *Interspeech 2011*, pages 2029–2032.

Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., e Lee, G. (2018). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. *arXiv preprint arXiv:1810.12541*.

Zhou, Y., Barnes, C., Lu, J., Yang, J., e Li, H. (2019). On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753.