# Concatenative Synthesis for Brazilian Sign Language

**Gustavo Borba da Silva**[1], **Ângelo Brandão Benetti**[1], **José Mario De Martino**[1]

[1]Avenida Albert Einstein, 400
Departamento de Engenharia de Computação e Automação
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas

g237965@dac.unicamp.br, angelobbenetti@gmail.com, martino@unicamp.br

***Abstract.*** *This paper presents an enhanced concatenative synthesis framework for Brazilian Sign Language (Libras) based on motion-capture (MoCap) data. The proposed method extends previous work by incorporating movement dynamics into the unit-selection process, allowing fragment selection to consider both pose similarity and velocity continuity. In addition, a quaternion-based transition smoothing mechanism is introduced to reduce discontinuities between independently recorded sign segments. The framework also employs a compact quaternion-only motion representation and a web-native architecture for real-time synthesis and visualization.*

## 1. Introduction

Sign language avatars provide a means of presenting signed content in digital environments, supporting accessibility in education, public services, and web communication. Generating movements that faithfully reproduce the dynamics of natural signing remains challenging: current machine learning and generative AI approaches lack the fine-tuning necessary for the language's nuances, while traditional concatenation-based methods often exhibit discontinuities in posture, velocity, or trajectory, producing unnatural animations that may reduce intelligibility and user acceptance among fluent signers.

Our concatenative synthesis offers an alternative approach. Originally developed for speech synthesis, concatenative methods generate new utterances by selecting and concatenating segments from a corpus of pre-recorded data [Hunt and Black 1996, Schwarz 2005]. Applied to sign language, this paradigm replaces speech units with motion-captured (MoCap) sign fragments: given an input gloss sequence, the system retrieves candidate instances from a MoCap database and combines them into a previously unseen animation. Because the output is constructed from authentic human performances, concatenative synthesis can preserve subtle kinematic and temporal characteristics that are often difficult to reproduce using procedural, neural, or manual animation techniques.

The feasibility of this approach for Brazilian Sign Language (Libras) using a gloss-annotated MoCap database is demonstrated in [Poeta 2025, Martino and Poeta 2021]. However, selecting fragments that produce smooth transitions remains a significant challenge. Existing methods rely primarily on pose similarity, which can reduce spatial discontinuities but does not account for movement dynamics, frequently resulting in abrupt velocity changes. Moreover, coarticulation between

consecutive signs makes segmentation and transition synthesis inherently challenging [Naert et al. 2017].

To address this issue, the paper presents an enhanced concatenative synthesis framework for Libras incorporating movement dynamics into unit selection, an improved quaternion-based transition smoothing mechanism considering angular momentum, and a compact quaternion-only motion representation. The framework targets a web-native architecture for real-time sign synthesis and visualization.

## 2. Methodology

### 2.1. Motion Capture Acquisition and Annotation

Motion capture sessions were conducted with fluent Brazilian Sign Language (Libras) signers using a wearable full-body motion capture suit. During each recording session, participants performed complete sentences while the system continuously tracked the orientation and movement of the body at a fixed sampling rate. This process generated a time-synchronized sequence of skeletal poses representing the kinematics of the signing performance.

Facial expressions and non-manual markers were recorded simultaneously using a head-mounted camera rigidly attached to the signer. The recorded video was subsequently processed using computer vision techniques to extract facial features and animation parameters associated with relevant facial movements. Synchronization between the body skeleton and facial recordings is key to ensure that manual and non-manual components of the signing performance can be reconstructed consistently. Currently, the synchronization approach is under development.

After acquisition, the captured motion was retargeted to the skeleton of the virtual avatar, establishing a correspondence between the performer's movements and the avatar rig. The resulting animation data were then converted into a compact representation consisting exclusively of joint rotations expressed as quaternions. By storing only quaternion values for each skeletal joint and frame, the system significantly reduces storage requirements while preserving the information necessary for accurate animation reconstruction and real-time playback.

To establish a correspondence between MoCap recordings and linguistic units, each recording session is rendered as a synchronized avatar video and annotated using the ELAN linguistic annotation tool [Max Planck Institute for Psycholinguistics 2025]. Fluent Libras interpreters identify the temporal boundaries of each sign by marking its start and end frames and assigning the appropriate gloss labels. Determining precise sign boundaries is difficult because signing is continuous and affected by coarticulation between neighboring signs [Naert et al. 2017]. The resulting annotations are synchronized with the underlying MoCap data, allowing the automatic extraction of sign-specific motion segments from continuous signing sequences. Each extracted segment is then associated with its gloss and stored in a motion database, enabling efficient retrieval and reuse during the concatenative synthesis process.

Our corpus focuses on three different lexical domains: health - 4,341 sentences (400 already recorded), food - 1023 recorded sentences, and transportation 1,321 recorded

sentences. Transportation will be recorded twice to cover standing and seated passenger and driver interaction.

### 2.2. Concatenative Synthesis Pipeline

The proposed concatenative synthesis pipeline consists of four stages: unit selection, content generation, transition smoothing, and animation synthesis. Starting from a sequence of glosses, the system retrieves motion fragments from the database, selects the most compatible instances, generates smooth transitions between consecutive fragments, and finally renders the resulting animation on a three-dimensional signing avatar.

#### 2.2.1. Unit Selection

For each gloss in the input sequence, all corresponding sign instances stored in the motion database are retrieved. Following [Poeta 2025], the selection problem is formulated as a shortest-path search on a graph in which nodes represent candidate sign instances, and edges represent transition costs [Schwarz 2005].

Unlike previous approaches, the proposed method considers both pose similarity and movement dynamics. In addition to the quaternion difference between consecutive fragments, the cost function incorporates velocity differences, thereby favoring transitions that better preserve movement continuity.

#### 2.2.2. Content Generation

Once the optimal fragment sequence has been identified, the corresponding quaternion trajectories are retrieved from the database and assembled into a continuous motion sequence. This sequence serves as the input to the transition smoothing stage.

#### 2.2.3. Transition Smoothing

To reduce discontinuities between independently recorded fragments, intermediate frames are generated using Spherical Linear Interpolation (SLERP) [Shoemake 1985], an approach proven effective for smoothing joint rotation transitions between discrete sign language motion units that doesn't jeopardize precision in order to exhibit smoothness [Song et al. 2012]. Moreover, a dynamic adjustment factor based on local velocity differences is additionally employed by preserving angular momentum to improve temporal continuity while maintaining the expressive characteristics of the original performance.

#### 2.2.4. Animation Synthesis

The final stage applies the synthesized motion sequence to a three-dimensional avatar in real time. Quaternion values are continuously mapped to the avatar skeleton, producing the final Libras animation at interactive frame rates for visualization in a web environment.

## 3. System Architeture

The proposed framework was designed with web compatibility as the primary requirement. The system was implemented as a TypeScript application using Three.js, an open-source framework for real-time three-dimensional rendering in web browsers. Future versions are expected to support standalone desktop deployment through Tauri, a lightweight Rust-based application framework.

Figure 1 illustrates the overall architecture of the proposed system. Initially, a user selects a textual input through a web interface. The text is then translated into a sequence of glosses by an external text-to-gloss module. Subsequently, the concatenative synthesis engine retrieves the corresponding motion fragments from the motion database, generates the final animation sequence, and renders it in real time using a three-dimensional signing avatar.

The synthesized animations are represented exclusively by quaternion trajectories, enabling efficient transmission and rendering while preserving motion quality.

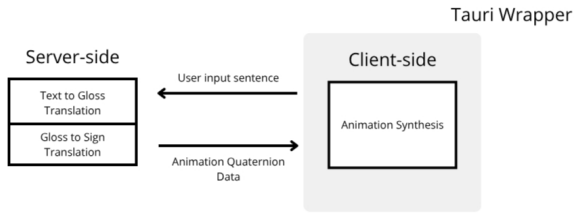


**Figura 1. Architecture of the Synthesizer (Source: The Author)**

## 4. Evaluation Methodology

The proposed framework will be evaluated through qualitative and quantitative analyses.

The qualitative evaluation will involve fluent Libras users who will assess synthesized animations using a five-point Likert scale. Evaluation criteria include naturalness, transition quality, hand articulation, facial expressiveness, intelligibility, and overall quality. The exact protocol which will be used is still under development.

The quantitative evaluation will compare synthesized transitions with corresponding transitions extracted from original motion-capture recordings. Similarity will be measured using Mean Squared Error (MSE) and the coefficient of determination ($R^2$) computed over quaternion trajectories. These metrics will provide an objective assessment of the fidelity of the synthesized motion.

## 5. Concluding Remarks

This paper presented an enhanced concatenative synthesis framework for generating Libras animations from motion-capture data. By incorporating movement dynamics into the unit-selection process and in the transition smoothing step, the proposed approach seeks to improve motion continuity and reduce visual discontinuities between concatenated sign fragments.

The framework combines a gloss-annotated motion database, a compact quaternion representation, and a web-native rendering architecture, enabling real-time synthesis and visualization of sign language animations.

Future work will focus on the seamless integration of acquired facial expressions and non-manual markers in the concatenative synthesis framework, as well as the investigation of learned motion representations based on self-supervised learning techniques to further improve fragment selection and transition quality.

## 6. Acknowledgements

This work is funded by the São Paulo Research Foundation (FAPESP), Grant No. 2026/05600-6 (Undergraduate Research Scholarship), Grant No. 2024/18575-4 (Technical Training Level IV), and Grant No. 2024/00914-7 (Science for Development Center: Assistive Technology and Accessibility in Brazilian Sign Language – CCD-TAAL).
Generative AI was utilized exclusively to improve the language, readability, and structural cohesion of this paper.

## Referências

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 373–376 vol. 1.

Martino, J. M. D. and Poeta, E. T. (2021). Método para a síntese concatenativa de línguas de sinais para a geração de avatares sinalizadores tridimensionais realistas. Patente Depósito: BR 10 2019 028311-4; Publicação: BR102019028311A2.

Max Planck Institute for Psycholinguistics (2025). Elan (version 7.1) [computer software]. Accessed: 2026-06-04.

Naert, L., Larboulette, C., and Gibet, S. (2017). Coarticulation analysis for sign language synthesis. In *Universal Access in Human-Computer Interaction. Designing Novel Interaction Experiences*, volume 10278 of *Lecture Notes in Computer Science*, pages 55–73. Springer, Cham.

Poeta, E. (2025). Síntese concatenativa de línguas sinais. Msc thesis, Universidade Estadual de Campinas(UNICAMP).

Schwarz, D. (2005). Current research in concatenative sound synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 1–4, Barcelona, Spain. International Computer Music Association.

Shoemake, K. (1985). Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '85)*, pages 245–254. ACM.

Song, D., Wang, X., and Xu, X. (2012). Chinese sign language synthesis system on mobile device. *Procedia Engineering*, 29:986–992.