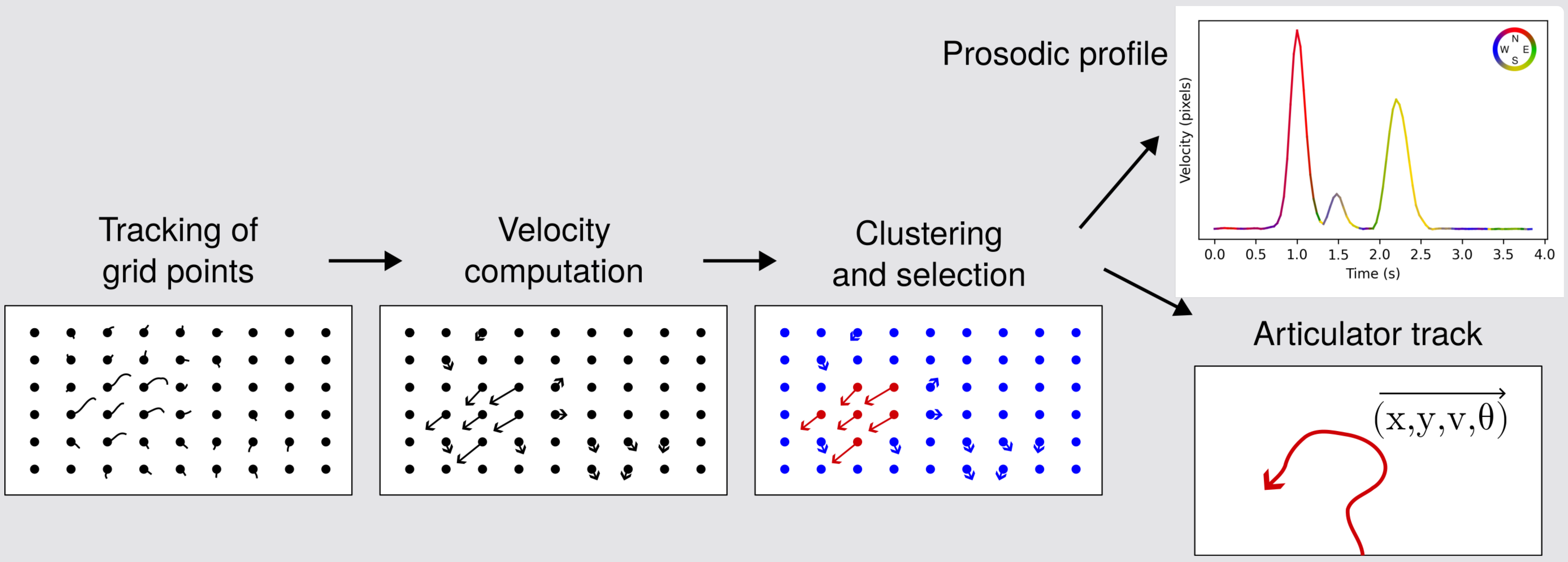
Automated Extraction of Prosodic Structure from Unannotated Sign Language Video

Antonio F. G. Sevilla
afgs@ucm.es

José María Lahoz-Bengoechea
jmlahoz@ucm.es

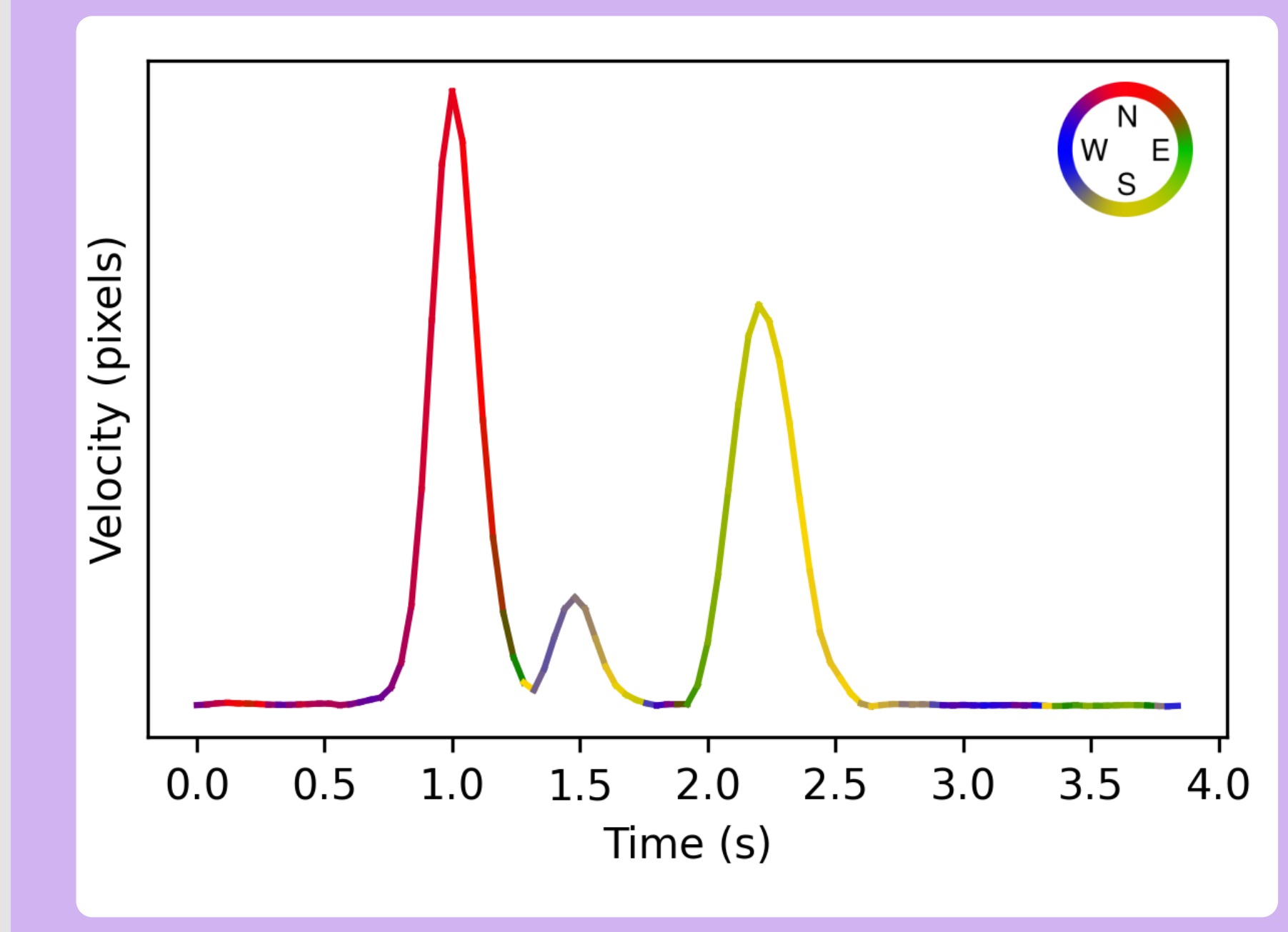
Alberto Díaz Esteban
adiazest@ucm.es



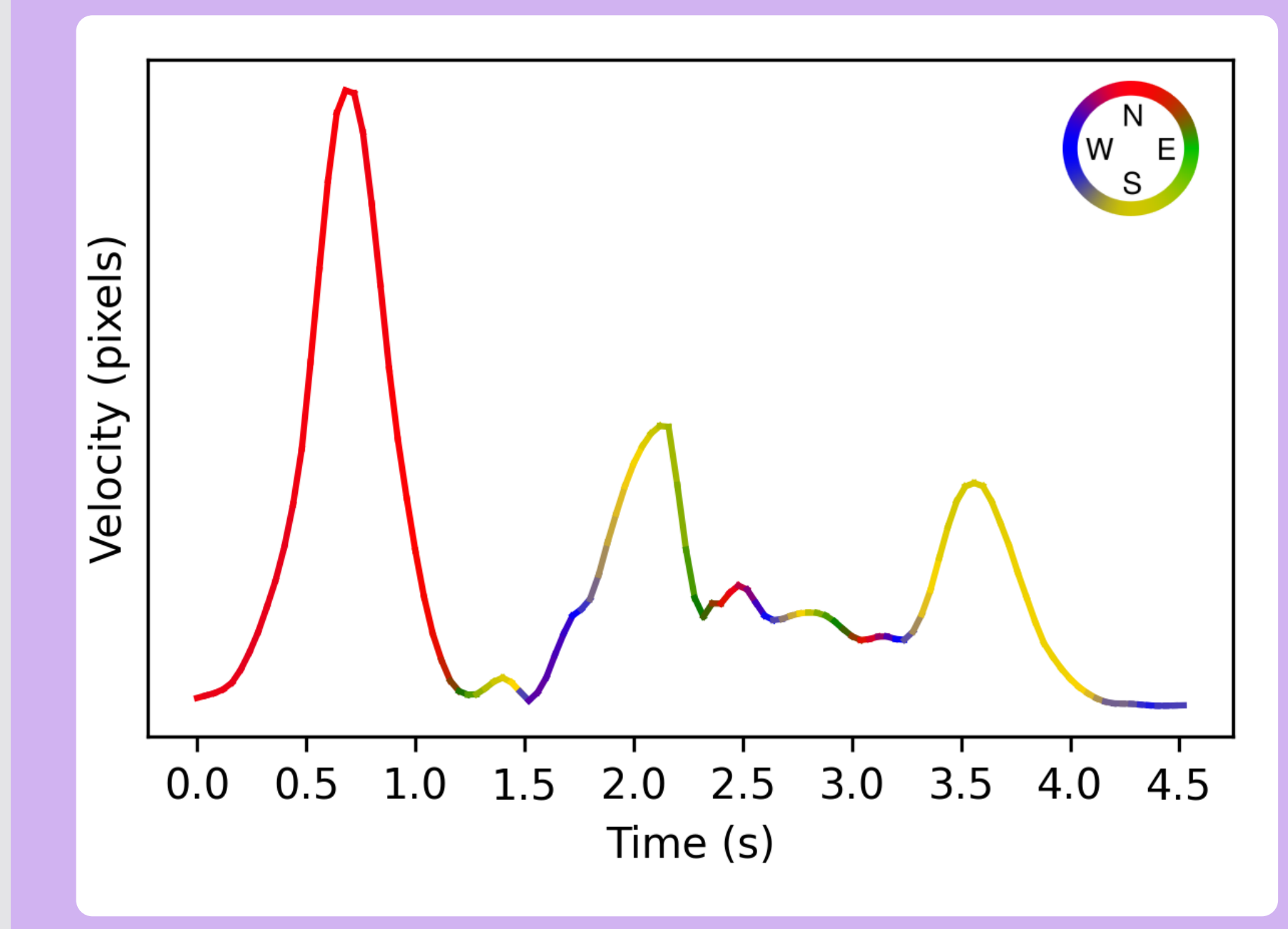
Prosody is an important carrier of linguistic information in sign languages. One prominent way this manifests is through the temporal structure of signs, including their **rhythm** and **intensity of articulation**. To empirically observe these effects, the velocity of the hands can be computed throughout the execution of a sign.

We propose a method for extracting this information from **unlabeled videos** of sign language, utilizing **CoTracker**, a recent advancement in computer vision that can track every point in a video without any calibration or fine-tuning. The dominant hand is identified through clustering of the computed point velocities, and its dynamic profile is plotted to reveal the prosodic structure of signing.

Prosodic Profiles



An MH sign: BSL "LOOK". Segments are clearly identifiable in the plot, and conform to Liddell and Johnson's model of movements and holds.



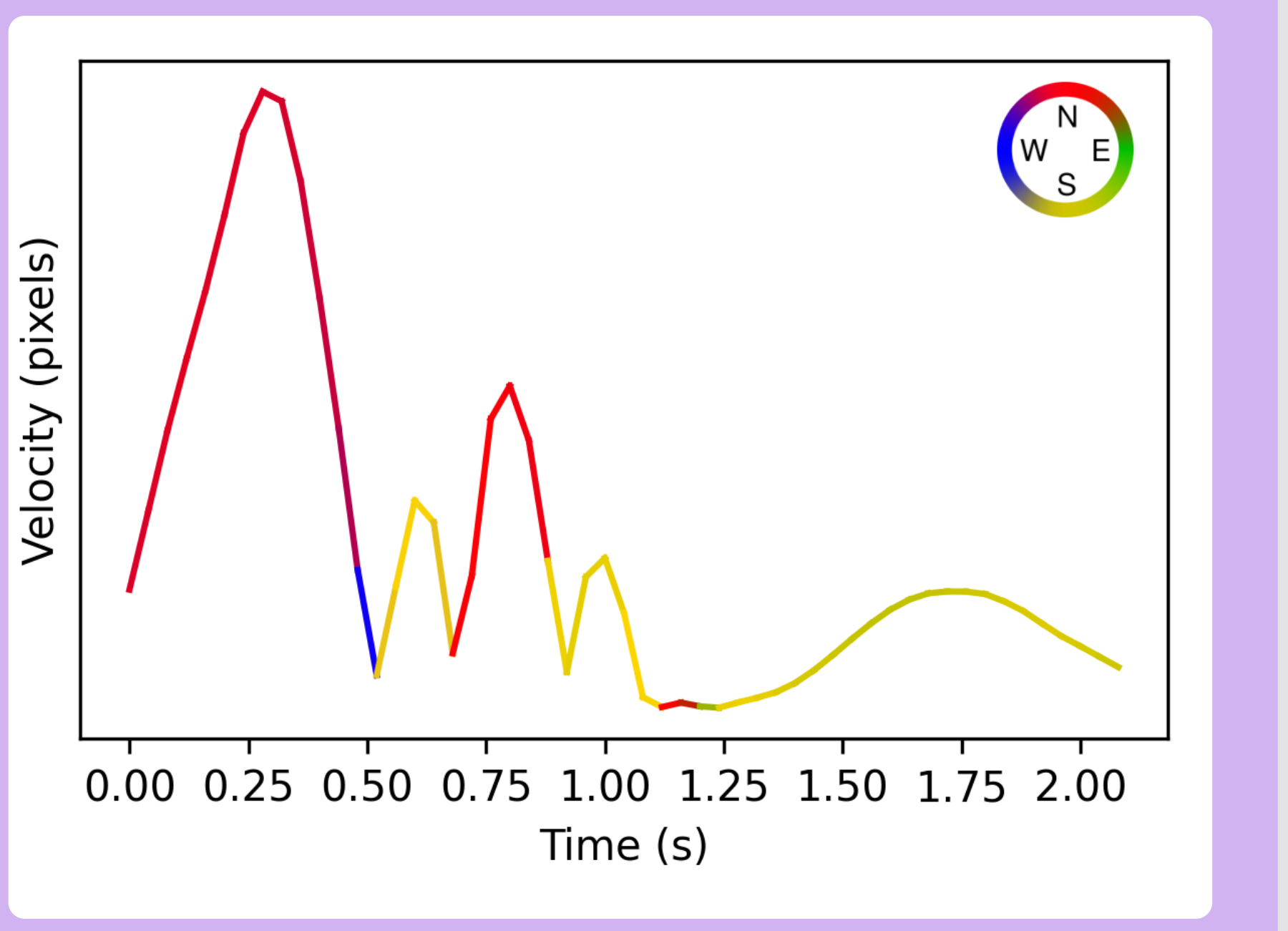
A compound sign in LSE: "FIREFIGHTER". There is a brief segment, which is either a hold or a small movement for contact, for the helment, and then a circular movement for the hose.

We plot the **velocity of articulation** (speed of the hand(s)) throughout the video duration. This generates a prosodic profile that helps elucidate the temporal structure of signing: distinct segments produce identifiable regions in the plots. Movement direction (line color), distinguishes changes of regime in articulation and aids in the identification of segments.

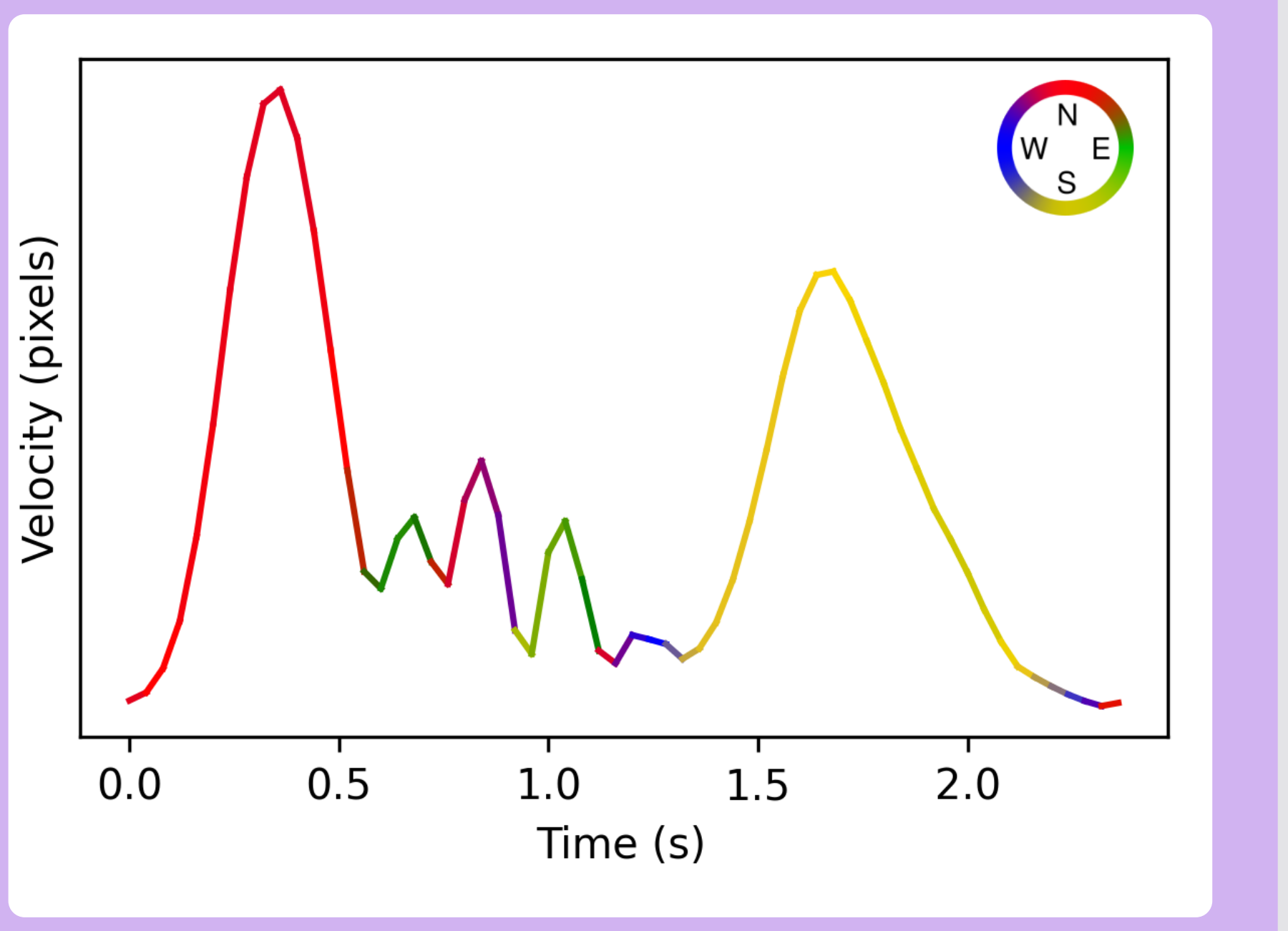
Movement segments (M) are characterized by peaks in velocity as the hand moves between target points, while **hold segments (H)** are plateaus with minimal movement, where the hand is held in position. Circular segments appear as regions of sustained elevated speed, but **slow enough to be perceptible**.

The segments with higher velocity are not necessarily the most significant. For instance, **preparation and accomodation** segments occur at the beggining and end of signing, or between components of a compound sign.

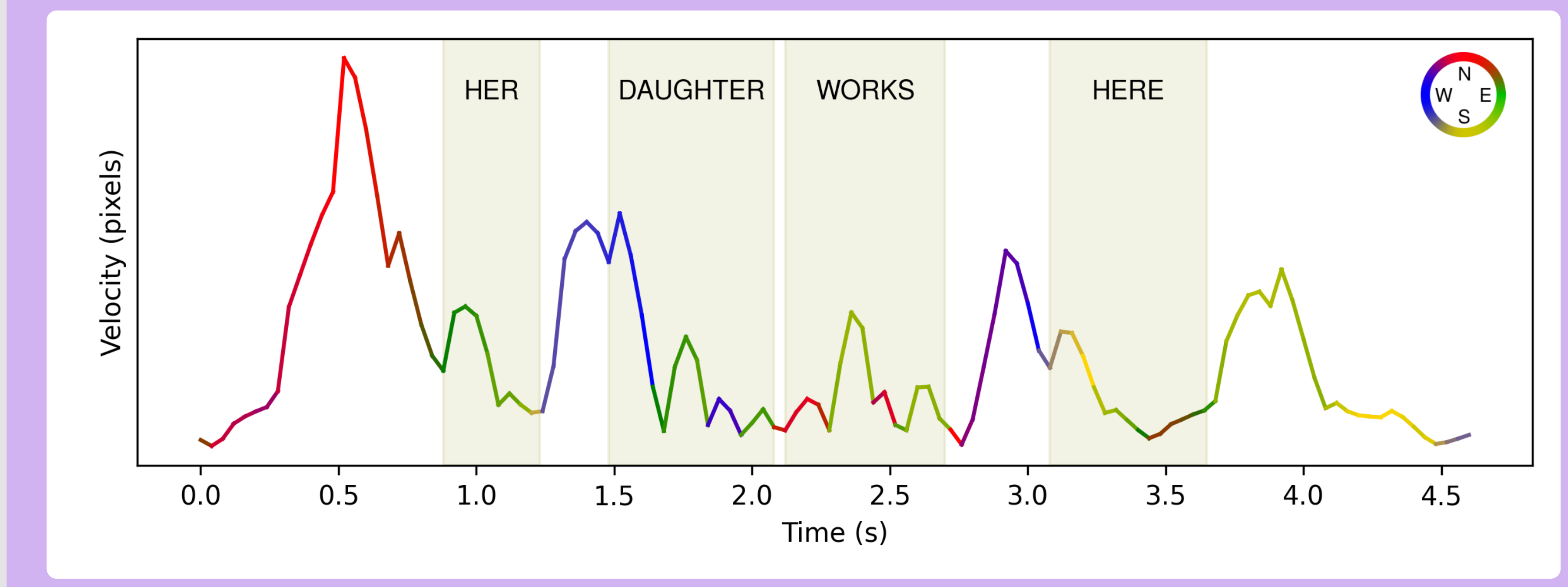
This pattern is also observed in signs with double repetition, where are typically three peaks: the two lexical strikes and an accomodation segment in between. Remarkably, this structure is highly consistent across different languages.



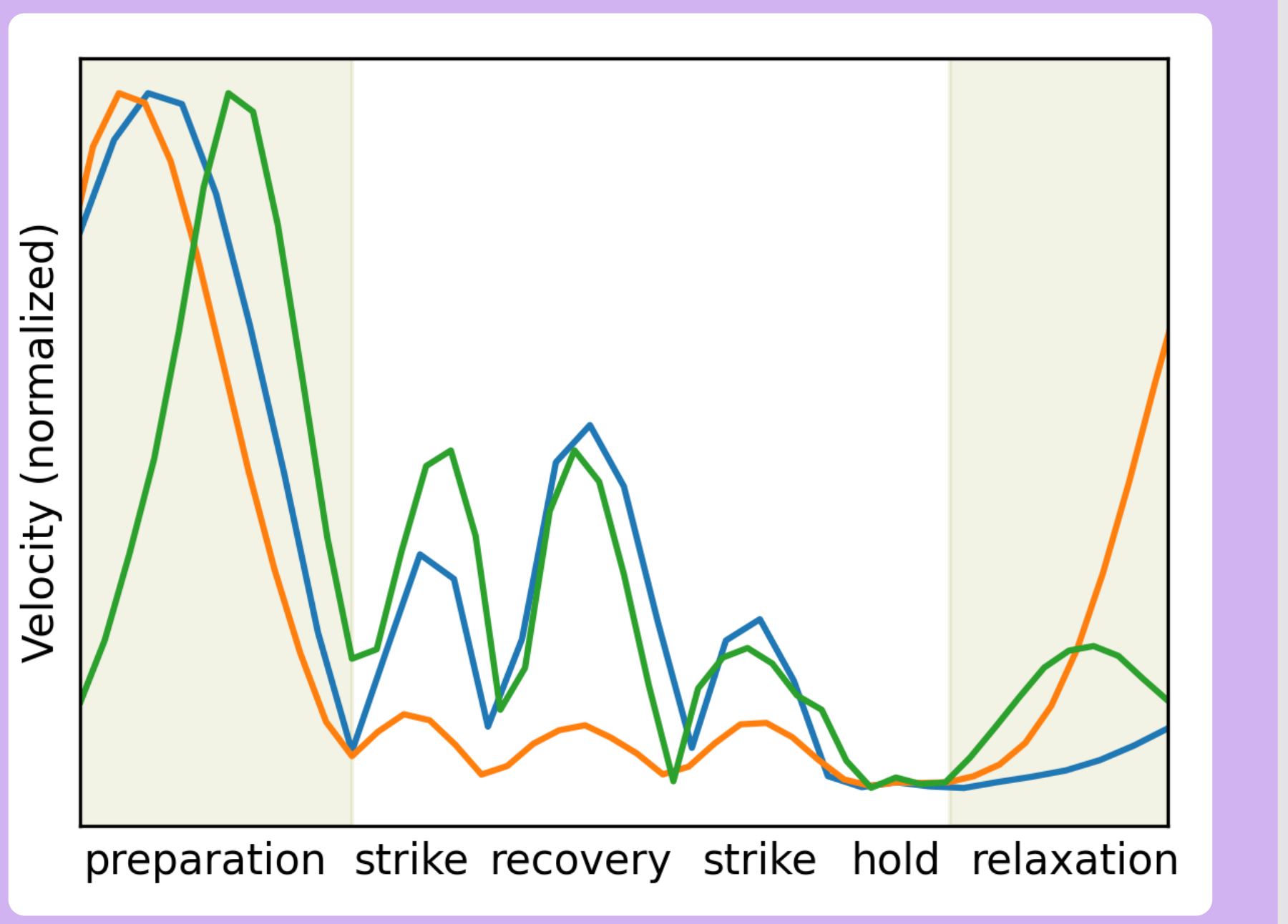
MM sign in LSE: "NEVER". Between the two lexical segments, an accomodation segment can be seen.



Another MM sign in LSE: "COUSIN". The often called "two syllables" here are similar to those in "NEVER", but not to the ones in "FIREFIGHTER".



BSL sentence: "HER DAUGHTER WORKS HERE", showing many features of prosodic structure in continuous speech.

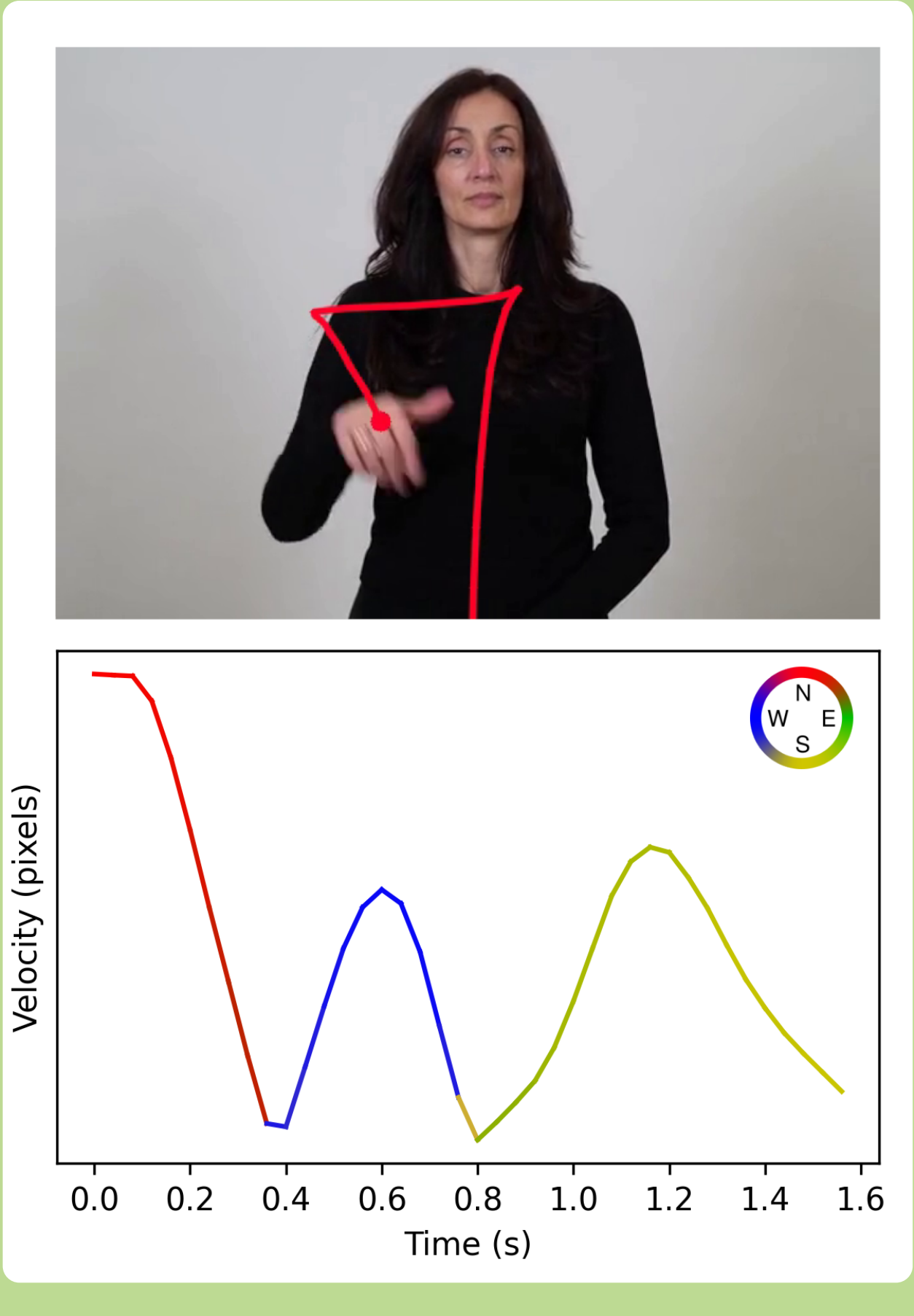


Similar MM structure can be observed across languages: ASL ("FROG"), BSL ("AUCTION") and LSE ("NEVER").

Our code is open source.
Try it on your own corpus!

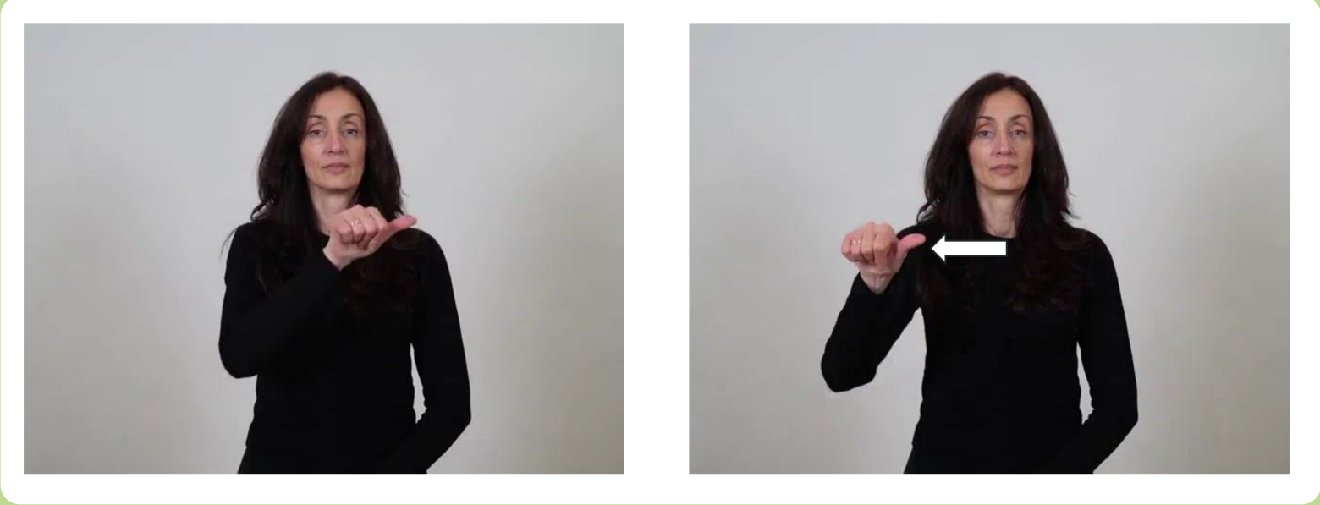
github.com/**agarsev/sign-prosody-extraction**

Articulator track



The articulator (hand) position can also be tracked in the video, shown here along with the corresponding prosodic profile.

Target points



Minima in the prosodic profile are key points in articulation, the "targets" of movement segments. We can extract thumbnails from the video at those points to present a static summary of sign articulation.

https://griffos.filol.ucm.es/**signario**
https://**bslsignbank**.ucl.ac.uk
https://**aslsignbank**.haskins.yale.edu
https://www.**spreadthesign**.com

Kopf, M., Schulder, M., and Hanke, T. (2022). "The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources"

Karaev, N., Rocco, I., Graham, B., Neverova N., Vedaldi A., and Rupprecht C. (2023). "CoTracker: It is Better to Track Together" github.com/**facebookresearch/co-tracker**

LREC-COLING 2024 — Torino (Italia), May 20-25 2024 — [HTML version] — [PDF version] — [Full text article]