Automated Extraction of Prosodic Structure from Unannotated Sign Language Video

Antonio F. G. Sevilla[1,2], **José María Lahoz-Bengoechea**[2], **Alberto Díaz Esteban**[1,3]
[1]Software Engineering and Artificial Intelligence, José García Santesmases, 9 28040 Madrid
[2]Spanish Linguistics and Literary Theory, Facultad de Filología edificio D 28040 Madrid
[3]Knowledge Engineering Institute, Facultad de Psicología, Lateral 2 28223 Pozuelo de Alarcón
Universidad Complutense de Madrid, Spain
afgs@ucm.es, jmlahoz@ucm.es, albertodiaz@fdi.ucm.es

Abstract

As in oral phonology, prosody is an important carrier of linguistic information in sign languages. One of the most prominent ways this reveals itself is in the time structure of signs: their rhythm and intensity of articulation. To be able to empirically see these effects, the velocity of the hands can be computed throughout the execution of a sign. In this article, we propose a method for extracting this information from unlabeled videos of sign language, exploiting CoTracker, a recent advancement in computer vision which can track every point in a video without the need of any calibration or fine-tuning. The dominant hand is identified via clustering of the computed point velocities, and its dynamic profile plotted to make apparent the prosodic structure of signing. We apply our method to different datasets and sign languages, and perform a preliminary visual exploration of results. This exploration supports the usefulness of our methodology for linguistic analysis, though issues to be tackled remain, such as bi-manual signs and a formal and numerical evaluation of accuracy. Nonetheless, the absence of any preprocessing requirements may make it useful for other researchers and datasets.

Keywords: Sign Language, Hand Tracking, Prosody

1. Introduction

The study of prosody and suprasegmental features in sign languages holds pivotal importance for our comprehensive understanding of their linguistic characteristics. These features, such as repetition and rhythm, serve as critical morpho-phonological parameters that contribute to many functions including, but not limited to, compound formation (Hwangbo and Choi, 2022), clause-level syntactic relationships (Malaia et al., 2013), and verbal inflection (Herrero Blanco, 2009). While the function of prosody is analogous to that in oral languages, its manifestation in sign languages is distinctly unique, largely owing to the pivotal role played by the motion of the articulators (Fenlon and Brentari, 2021).

The displacement and velocity profile of hand movements during signing has been posited as a potentially sufficient metric for analysing their prosodic structure, both at the lexical and the post-lexical levels (Wilbur and Martinez, 2002). The changes in velocity and direction can demarcate segment changes in phonological structure, but to empirically validate these theoretical postulates and to leverage them for rigorous linguistic analysis, there is a need for relevant data. In the realm of sign languages, these data often come in the form of videos, which do not inherently encode the linguistic parameters of interest.

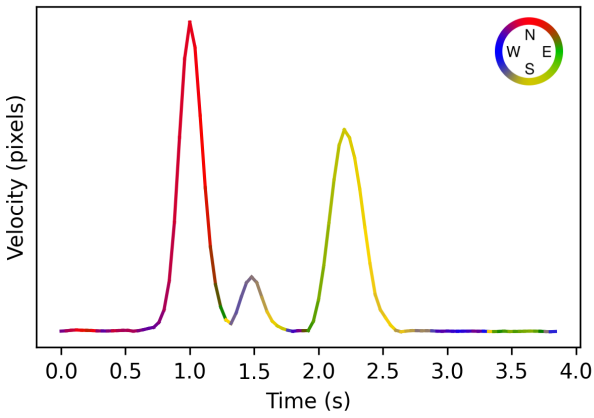

Figure 1: Dynamic profile of the BSL (British Sign Language) sign “LOOK”[1].

Recent advancements in computer vision have emerged as a solution to this challenge. CoTracker is a state-of-the-art deep learning architecture that allows for the tracking of arbitrary points in source videos, without requiring additional annotations or dataset-specific training (Karaev et al., 2023). This technology provides a pathway for computing and analyzing hand velocity using computational tools.

In our proposed method[2], we leverage the capabilities of CoTracker to compute the velocity of hand

[1]https://bslsignbank.ucl.ac.uk/dictionary/words/look-1.html

[2]https://github.com/agarsev/sign-prosody-extraction

LREC-COLING 2024, pages 1808–1816

movements in various sign languages. The subsequent analysis yields temporal plots that show the velocity and directional changes over time, enabling the visual and empirical examination of the signs' prosodic structure. Predominantly, the y-axis in the plots in this article represents the articulation velocity, measured in units of pixels per frame. Given that the absolute numerical values of velocity are neither pertinent nor informative for our analysis, they have been intentionally omitted. On the x-axis, time is delineated in seconds, corresponding to the duration of the video clip under investigation.

To enhance interpretability, we have opted to color-code the lines in accordance with the direction of articulator movement within the video frame. It is crucial to clarify that this is a two-dimensional representation of direction as perceived in the video, and while it is related, it is not synonymous with the three-dimensional direction in the actual sign-space, especially when movements in the back and forth axis are involved. To demarcate this distinction, we use the cardinal directions (N, E, S, W).

For example, in Figure 1, we see a BSL (British Sign Language) sign consisting of a single movement. The graph shows three corresponding velocity peaks: 1) an upward preparation; 2) the primary forward motion; and 3) a downward relaxation. Notably, a brief hold is present after the primary motion, but absent at the onset.

In Figure 2, we can see two signs from LSE (Spanish Sign Language). The profile of these two signs reveals they have the same class of prosodic structure. They exhibit two primary strikes, along with a preparatory movement for repetition between them in the opposite direction. A brief hold is observed at the termination of both signs, though this may be potentially attributable to the specific intonation profile associated with signs intended for dictionary inclusion.

These visual representations serve a dual purpose: they elucidate both the segmental and suprasegmental structures that are integral to the articulation of signs. When interpreting these plots, it is important to recognize that the most prominent peaks—those indicating higher velocities—are often not indicative of lexical segments within signs. Rather, they frequently correspond to transitional or accommodative hand movements that occur as the hand navigates between distinct spatial locations. For instance, the initial peak in each plot typically signifies preparatory movements, transitioning the hand from a resting position to the first signing location.

While the primary emphasis of this article is on the methodological framework, our preliminary findings offer compelling observations and avenues for discussion. These results have been generated from multiple datasets encompassing a range of

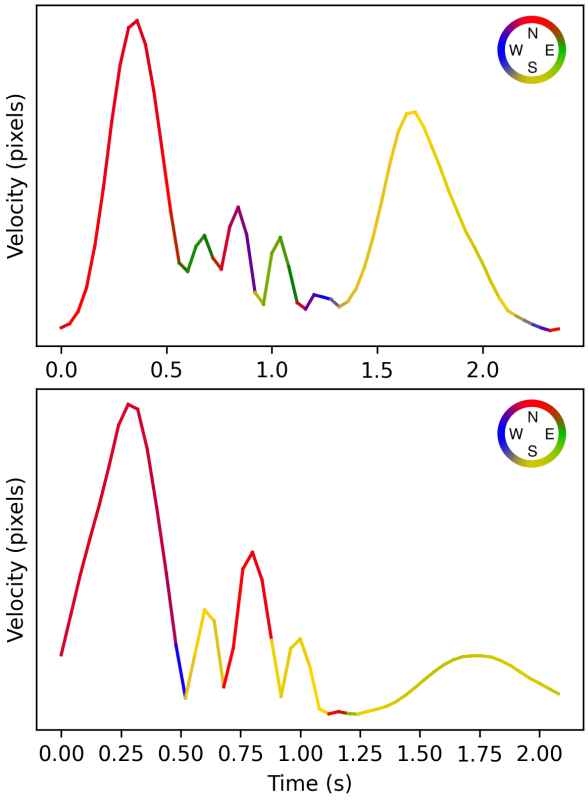

Figure 2: Dynamic profile of LSE signs "COUSIN"[3] and "NEVER"[4].

sign languages, thereby suggesting the potential for a more universal applicability of our methodology. Such a universal scope could hold significant implications, extending beyond the realm of linguistic analyses to encompass broader computational processing tasks.

The remainder of this article is organized as follows: section 2 describes previous methodologies that have been employed for similar purposes, while section 3 explicates our proposed methodology. Some preliminary findings that signify the promise of our approach are explored in section 4, and section 5 draws some conclusions from them. Finally, section 6 offers a reflective analysis of the limitations inherent in our proposal, along with potential avenues for future research.

2. Background

As outlined in the introduction, the prosodic structure of sign language plays a crucial role in its interpretation. Prosody encompasses the hierarchical organization of individual segments into more

[3] https://griffos.filol.ucm.es/signario/signo/13227

[4] https://griffos.filol.ucm.es/signario/signo/11169

complex structures, focusing among others on elements of timing and rhythm. Analogous to its significance in oral languages, prosody serves as an essential component for conveying meaning in sign languages. Conducting empirical research in this domain necessitates accurate measurements and quantitative data. One plausible approach to acquire these data is through the extraction of kinematic information, specifically focusing on articulator displacement and its associated attributes—duration and velocity. Wilbur and Martinez (2002) posits that such information sufficiently encapsulates the dynamic structure of sign language.

Various methodologies have been employed to achieve this aim. For instance, Borneman et al. (2018) utilized complexity analysis of optical flow to estimate velocity parameters. However, this approach offers a global measurement and fails to discriminate between the movements of different hands or directions. To obtain hand trajectories, Wilbur and Zelaznik (1997) employed a 3D motion analyzer system, Koech (2007) used a combination of electromagnetic sensors and wearable gloves, and Abdullahi and Chamnongthai (2022) sourced data from a Leap Motion Controller. While effective, these techniques entail complex experimental setups and specialized datasets, rendering them inaccessible to broad research applications.

Alternatively, 2D video datasets, such as those cataloged in Chapter 5 of the Sign Language Dataset Compendium (Kopf et al., 2022), offer a more ubiquitous resource. While these datasets sometimes feature linguistic annotations, they often lack the specific timing or kinematic data required for this research. Although manual annotation is possible via tools like ELAN[5], it is typically geared towards marking independent signs in discourse, like in Crasborn et al. (2016), rather than detailing the internal structure of signs.

Automatic extraction presents a more efficient alternative and makes utilization of existing 2D video corpora feasible. One approach is to estimate 3D hand or body pose from 2D videos. This can entail the complexities of annotation and dataset specificity (Ohkawa et al., 2023), although recent developments using machine learning show great promise in mitigating these issues (Börstell, 2023). Another option involves 2D object tracking, usually carried out in two steps: hand detection and subsequent tracking across video frames (Yuan et al., 2005). However, the variability in video settings and anatomical differences present challenges to hand detection (Thangali and Sclaroff, 2009). Recently developed tracking technologies overcome these limitations by tracking points within source videos without requiring prior object information (Neoral et al., 2023; Wang et al., 2023; Karaev et al., 2023).

While promising, these point-tracking methods do not directly provide information on hand movements or velocities. They yield traces of points that require further analysis. Mcdonald et al. (2016) offer methodologies for analyzing such position and velocity series in the context of prosody, albeit within avatar generation. To alleviate the problem of noise, and as precise computing of velocity minima is crucial for identifying articulatory changes, Savitzky-Golay filters (Savitzky and Golay, 1964) can be employed to smooth the series and its derivatives.

The final unresolved issue is the identification of hands based on point velocities. Given that we are working with videos focused on sign language, it is reasonable to assume that the fastest-moving objects will be the hands. K-means clustering (Lloyd, 1982) can segregate the point traces into separate classes based on velocity, effectively isolating the articulators from the largely static background.

In the subsequent section, these individual components will be synthesized into a cohesive pipeline capable of automatically extracting velocity plots from unannotated sign language videos.

3. Methodology

To develop our methodology, and for testing and visualization in this article, we have used videos from various sources, including the ASL (American Sign Language Hochgesang et al., 2023) and BSL (British Sign Language Fenlon et al., 2014) Sign Banks, and dictionaries such as SpreadTheSign (Hilzensauer and Krammer, 2015), and the Spanish Sign Language Signary[6]. These videos are then processed in a series of steps, schematically represented in Figure 3 and detailed in the following. Our code is also freely available online at GitHub[7].

3.1. Tracking of Grid Points

The initial step in our pipeline involves utilizing CoTracker (Karaev et al., 2023) to track distinct points within the video footage. To minimize computational cost, a uniform grid of 30x30 points across the frame is selected for tracking, rather than every individual pixel. Tracking commences at the midpoint of each video, to ensure that the hands are actually in the frame. Tracking is then conducted both forwards and backwards, with the latter being reversed and then prepended to the forward tracks, ultimately yielding a set of 900 tracks that trace different points throughout the entire video.

Opting for a uniform grid spanning the full frame circumvents the need for detection or manual an-

[5] https://archive.mpi.nl/tla/elan

[6] https://griffos.filol.ucm.es/signario

[7] https://github.com/agarsev/sign-prosody-extraction

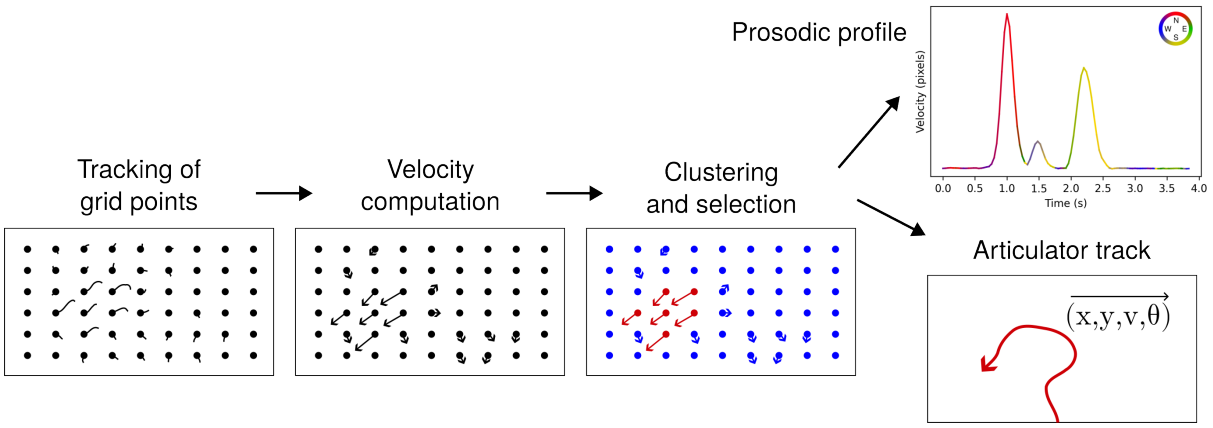

Figure 3: Steps of the processing pipeline.

notation of hand location at any given time point. Moreover, this approach seems to aid the method not to get confused in the presence of overlapping body parts that share similar skin tones.

The tracking process is the most computationally intensive step, but can be executed on reduced-quality videos without compromising the reliability of the results. Utilizing high-end but consumer-grade hardware, an NVIDIA GeForce RTX 3080 GPU (Ampere architecture) and Intel Core i9-9900K CPU, each video takes approximately 10 seconds to process, varying with video length.

3.2. Velocity Computation

The output from CoTracker is generated in PyTorch[8] format, which can be easily imported into Python scripts for further processing. Each predicted trace from CoTracker consists of a sequence of x and y coordinates, representing each tracked point, at every frame. Spatial coordinates are differentiated along the temporal axis to obtain velocities in x and y. These velocities are then converted to polar coordinates, to isolate absolute velocity and movement direction. While our primary focus lies on absolute velocity for the extraction of prosodic features, the directional of the movement can aid in visualization and segment discrimination. To compute the time derivative, the Savitzky-Golay filter as implemented in the SciPy library is employed (Luo et al., 2005; Virtanen et al., 2020).

3.3. Clustering and Selection

The subsequent step involves separating the tracks into two groups, based on their velocities. This way, we can separate the tracks into those corresponding to the articulator, and those corresponding to the background and body. We use K-means clustering (Lloyd, 1982), in particular scikit-learn's implementation of the algorithm (Pedregosa et al., 2011). The cluster with higher velocity is interpreted as the main articulator, and its center, the average of all the tracks that comprise it, is then used as the final result for the articulator's velocity and direction of movement.

3.4. Prosodic Profile Extraction

Velocity plots are then generated using Matplotlib (Hunter, 2007), where velocity is plotted on the y-axis against each video frame on the x-axis. Peaks and valleys in these plots reveal the underlying prosodic structure. The lines in the plot are colored according to the direction of movement, which helps in segment identification and correlation with lexical components of the sign. Examples of this have been shown in Figures 1 and 2.

To quantify these observed segments, we compute local minima and maxima of the velocity curve. These extrema are calculated by differentiating the velocity using the Savitzky-Golay filter and identifying zero-crossings. Local minima correspond to moments of slowed hand movement, which we propose separate the segments in the framework for sign language description proposed by Liddell and Johnson (1989). Local maxima are utilized to fine-tune the detection of key points and to eliminate extraneous points that lie outside the bounds of the articulated sign.

3.5. Articulator Location Recovery

The focus of our methodology lies in extracting the velocity series in order to plot the dynamic profile of the sign. This profile contains information on segment duration and intensity, which we believe provide a very good characterization of its prosodic structure. Nevertheless, spatial displacement information can hold additional utility.

Upon isolating the principal articulator based on velocity, we can extract the original x and y coordi-

[8] https://pytorch.org/

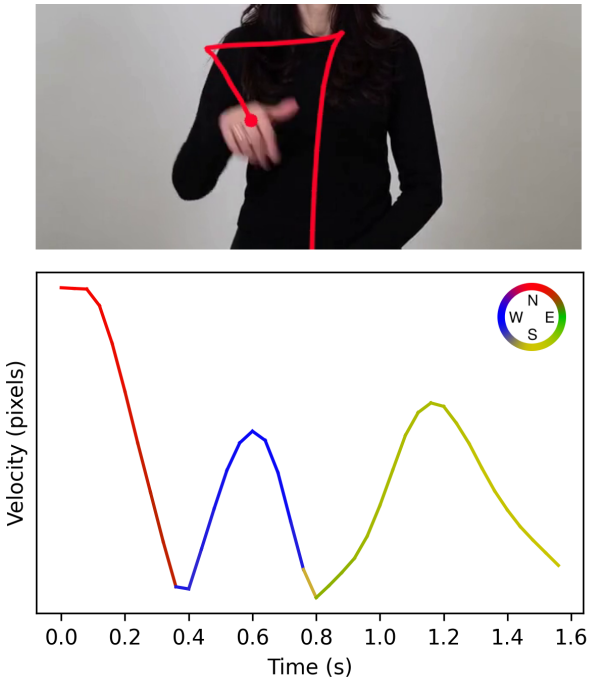

Figure 4: Path of movement along velocity profile for the LSE sign "WEEK"[9].

nate tracks for each point in the cluster. The arithmetic mean of these positional vectors pinpoints the articulator's location throughout the video sequence, as illustrated in Figure 4. There, we can see the trajectory of the hand superimposed in red on a still frame from the video. The corresponding velocity profile illustrates the distinct phases: an initial upward preparation, a brief pause at the onset, the principal movement directed westward (ipsilateral, as the subject is right-handed), and the subsequent relaxation phase. Although we primarily use this visualization to validate our methodology, this spatial information may also be useful for further analyses which require precise location data. However, pose estimation methodologies such as the one presented in Börstell (2023) may be better suited for this specific objective.

4. Exploratory Analysis

In previous figures we have explored the dynamic profiles of multiple simple signs. These plots show empirical visualizations of the different segments constituting each sign, facilitating an understanding of their sequential arrangement, relative duration, and velocity characteristics, but our method is not limited to such signs.

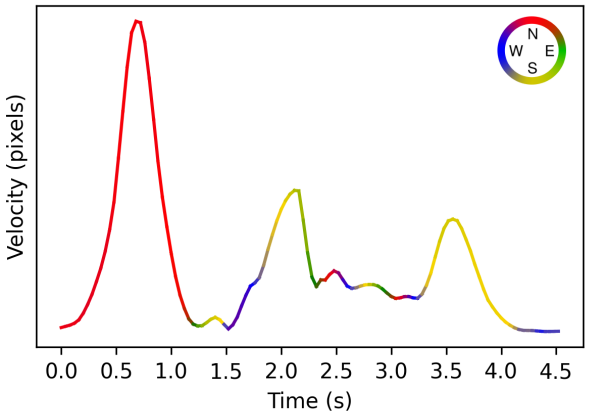

Figure 5: Dynamic profile of the LSE compound sign "FIREFIGHTER"[10].

4.1. Compound Signs

We can apply our pipeline of computations to further investigate the prosodic profile of compound signs, as exemplified in Figure 5. The first morpheme of this compound sign emulates a firefighter's helmet, by holding an iconic handshape touching against the forehead. Our method captures a minor peak during this "hold" phase, suggesting an initial preparation of the handshape near its designated locus prior to the execution of the touch. Next, the sign transitions into a high-velocity movement directed toward the locus of the second morpheme, the neutral space. Here, a circular movement is enacted, depicting the firehose. The initial arc of this circle—progressing southward and then eastward—is executed in tandem with the preparatory phase, while the subsequent circular motion maintains a relatively constant velocity, remains spatially confined, and exhibits a narrower trajectory.

As with other signs we have examined before, Figure 5 again reveals that the segments with the highest velocity are not necessarily the most linguistically relevant. One might assume that higher-velocity segments would be more salient from a kinematic perspective. However, the key factor is perceivability. Movements that are too fast may be more difficult to perceive, whereas the lexically pertinent segments often exhibit lower velocity, enhancing their perceptual saliency. For instance, even though the "hose" morpheme is categorized as an M segment (dynamic) and has a relatively higher velocity compared to the H segment (hold) representing the helmet, its velocity remains constrained in relation to the perceptually irrelevant accommodative segments.

[9] https://griffos.filol.ucm.es/signario/signo/11395

[10] https://griffos.filol.ucm.es/signario/signo/11197

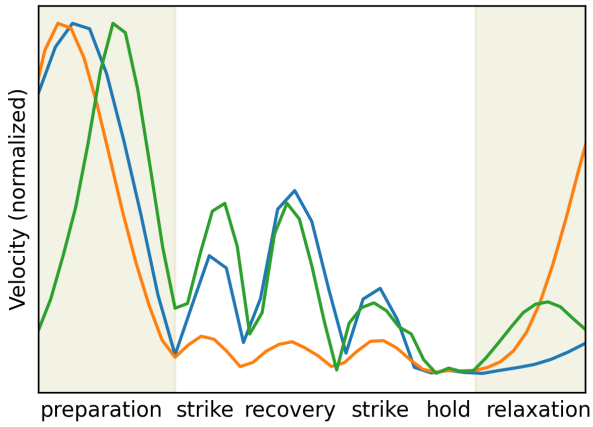

Figure 6: Phases in the articulation of some signs in ASL (FROG), BSL (AUCTION) and LSE (NEVER).

4.2. Comparative Analysis

In a similar vein to the analysis of different morphemes within a single sign, the juxtaposition of multiple signs offers another insightful avenue of exploration. To facilitate this comparison, one can superimpose various plots on a single graph. Given that velocity is not a consistent measure—especially when quantified in pixels per frame, a unit with limited interpretive value—we have elected to normalize the plots. Specifically, velocities are normalized to a range of 0-1, and time is re-scaled so that the lexical segment lies between 0 and 1.

Adopting this standardized framework allows for a more coherent comparison of the prosodic structures across different signs, as illustrated in Figure 6. In this figure, signs from three distinct sign languages are depicted. All of these signs belong to the class characterized by "repeated simple movements," and the underlying segmental structure is remarkably consistent across the languages. Furthermore, the duration of these segments also exhibits a notable degree of uniformity, underscoring the potential universality of certain prosodic elements in sign languages.

4.3. Sentence Level Analysis

In addition to lexical prosody, which pertains to the internal structure of individual signs, another salient aspect of sign language prosody occurs at the sentence level. This involves analyzing the rhythm and duration of various signs to infer elements such as phrasing. Our methodology is also amenable to analyses of longer videos that encompass complete sentences, provided certain conditions are met: the focus remains on a single signer, and the articulator is the fastest moving object in the frame. An example of this is presented in Figure 7, depicting a four-sign BSL sentence.

In this figure, both "DAUGHTER" and "WORKS" are characterized as two-strike signs, similar to the ones depicted in Figure 2. We can again see the distinct phases each sign undergoes: preparation, first strike, recovery, and second strike. The sign "HER" also consists of two strikes, but exhibits a different dynamic profile. We propose that this represents a distinct prosodic class of sign, specifically one with a movement-hold (MH) structure. In this case, the hold is not entirely static; rather, it incorporates minor motion to enhance its sonority. The sign "HERE" belongs to the same MH prosodic class, in this case characterized by an initial downward strike followed by an elongated hold, which probably is more pronounced due to the specific intonation pattern employed in this sentence.

It is noteworthy that the durations of "DAUGHTER" and "WORKS" are quite similar, although a more extensive corpus of examples is needed to substantiate any conclusive claims. Moreover, we can observe again how the locative constraints inherent in sign language execution significantly influence the preparation phases of the signs, often resulting in the most conspicuous peaks in the velocity curve.

5. Conclusions

Plots such as the one in Figure 7, along with the underlying numerical data, hold promise as invaluable tools for the linguistic and prosodic analysis of sign language. They provide insights into both the internal structure of individual signs and the rhythmic, velocity, and organizational attributes of full sentences.

A significant advantage of our methodology is its operational simplicity and accessibility. It circumvents the need for manual annotations and remains sufficiently robust for lower-quality video analysis. The algorithm is designed to run on consumer-grade hardware, eliminating the need for specialized computational resources, and enabling researchers to apply this methodology to their respective datasets. We anticipate that the synergistic combination of our approach and the growing availability of sign language data will contribute to an enhanced understanding of sign language prosodic structure.

Furthermore, part of our methodolgy involves the computation of a series of relevant points, local minima in velocity which can be seen as small pauses in articulation. We mainly use them to aid in the segmentation of signs for comparative analysis, but they can also be used to compute precise durations of segments, which may shed further light or provide empirical support to different theoretical models of sign language structure.

Beyond their immediate utility for segmentation,

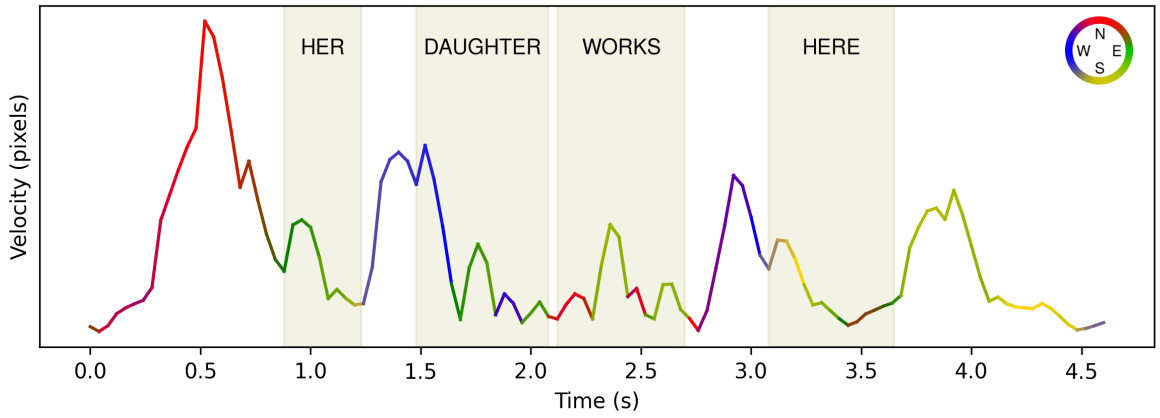

Figure 7: Dynamic profile of the BSL sentence "HER DAUGHTER WORKS HERE"[11].

the significance of these points may extend to further computational processes. By extracting the frames corresponding to these points, we can obtain static key-frames encapsulating essential phonological parameters. This feature can prove beneficial for the development of sign language dictionaries, as it allows for the automatic creation of thumbnails or static summaries that can be readily consulted in print form or other mediums where video is either unavailable or impractical.

6. Limitations and future work

The main limitation of our proposed methodology resides in its foundational assumption: it presupposes that the target video features a singular signing individual with no other dynamic entities present, thereby making the signing articulator the fastest-moving object in the frame.

Bimanual signs where both hands are active introduce a level of complexity, as our algorithm identifies both hands together as the articulator. While this often does not pose a problem—given that in these signs the velocity of the hands is usually synchronized—it remains a limitation.

Another inherent limitation stems from the use of 2D video, since movements in the antero-posterior plane may be visually diminished due to camera perspective. Our preliminary analysis suggests, however, that this limitation might be of minor concern, as movements are generally visible enough through their vertical or lateral components. For our research objectives, absolute measures of velocity are less critical than understanding the timing and sequencing of sign phases.

Nonetheless, these limitations underscore the necessity for future work to transform our informal evaluations into formal numerical assessments. We are in the process of systematically analyzing a corpus using our approach, where we hope to be able to extract meaningful statistics and linguistic conclusions.

Initially, we anticipated that such efforts would necessitate manual annotations, but thanks to a recent publication highlighted by one of our reviewers, we now see a promising alternative. Börstell (2023) uses recent advances in machine learning to automatically estimate body pose before conducting an analysis similar to ours. Future research could involve a comparative evaluation between both our approaches, and if findings from both methods align, it would strongly suggest the reliability of our measurements. Additionally, their technique could potentially refine the initial steps of our methodology, particularly in addressing challenges related to bimanual signs and precise articulator localization.

From a more linguistic point of view, additional research is also needed to examine continuous discourse and diverse settings, extending beyond our current focus on lexical items and prosodic elements. The dynamic profiling of longer, syntactically complex sentences promises valuable insights into the internal and sentence-level structure of sign language.

7. Acknowledgments

The research leading to and the publication of this article has been funded and made possible by the projects "Visualizando la SignoEscritura" (Visualizing SignWriting, https://www.ucm.es/visse), funded by Indra and Fundación Universia in the IV call for funding aid for research projects with application to the development of accessible technologies; "CANTOR: Automated Composition of Personal Narratives as an aid for Occupational Therapy based on Reminescence", Grant. No. PID2019-108927RB-I00 (Spanish Min-

[11] https://www.spreadthesign.com/en.gb/sentence/9976/her-daughter-works-here/

istry of Science and Innovation); and "Signario de LSE: Diccionario paramétrico de la lengua de signos española" (SSL Signary: A parametric dictionary of Spanish Sign Language, https://www.ucm.es/signariolse), reference number IN[21]_HMS_LIN_0070, supported by a 2021 Leonardo Grant for Researchers and Cultural Creators from the BBVA Foundation. The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

We want to acknowledge the collaboration of the signing community, especially the Spanish Sign Language teachers at Idiomas Complutense, and Fundación CNSE.

We thank the reviewers for their valuable suggestions, especially for pointing out relevant literature that has improved both this paper and our future research.

8. Bibliographical References

Abdullahi, S. B. and Chamnongthai, K. (2022). American Sign Language Words Recognition Using Spatio-Temporal Prosodic and Angle Features: A Sequential Learning Approach. *IEEE Access*, 10:15911–15923. Conference Name: IEEE Access.

Borneman, J. D., Malaia, E., and Wilbur, R. B. (2018). Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27(5):051229.

Börstell, C. (2023). Extracting sign language articulation from videos with MediaPipe. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 169–178. University of Tartu Library.

Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E., Schüller, A., Ormel, E., Nauta, E., van Zuilen, M., Winsum, F. v., and Ros, J. (2016). Linking lexical and corpus data for sign languages: Ngt signbank and the corpus ngt.

Fenlon, J. and Brentari, D. (2021). Prosody: Theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*.

Herrero Blanco, Á. (2009). *Gramática didáctica de la lengua de signos española (LSE)*. Ediciones SM.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Hwangbo, H. J. and Choi, Y. (2022). Morpho-phonological investigation of compounds in Korean Sign Language. *Studies in Phonetics, Phonology, and Morphology*, 28(1):169–198.

Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., and Rupprecht, C. (2023). CoTracker: It is better to track together.

Koech, C. (2007). A kinematic analysis of sign language. Master's thesis, New Jersey Institute of Technology.

Liddell, S. K. and Johnson, R. E. (1989). American Sign Language: The Phonological Base. *Sign Language Studies*, 1064(1):195–277.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Luo, J., Ying, K., and Bai, J. (2005). Savitzky–golay smoothing and differentiation filter for even number data. *Signal processing*, 85(7):1429–1434.

Malaia, E., Wilbur, R. B., and Milković, M. (2013). Kinematic Parameters of Signed Verbs. *Journal of Speech, Language, and Hearing Research*, 56(5):1677–1688.

Mcdonald, J., Wolfe, R., Wilbur, R., Moncrief, R., Malaia, E., Fujimoto, S., Baowidan, S., and Stec, J. (2016). A New Tool to Facilitate Prosodic Analysis of Motion Capture Data and a Data- Driven Technique for the Improvement of Avatar Motion.

Neoral, M., Šerých, J., and Matas, J. (2023). MFT: Long-term tracking of every pixel.

Ohkawa, T., Furuta, R., and Sato, Y. (2023). Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey. *International Journal of Computer Vision*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Thangali, A. and Sclaroff, S. (2009). An alignment based similarity measure for hand detection in cluttered sign language video. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 89–96. ISSN: 2160-7516.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, Q., Chang, Y.-Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., and Snavely, N. (2023). Tracking everything everywhere all at once. In *International Conference on Computer Vision*.

Wilbur, R. and Martinez, A. M. (2002). Physical correlates of prosodic structure in American Sign Language.

Wilbur, R. B. and Zelaznik, H. N. (1997). Kinematic correlates of stress and position in asl. In *Annual Meeting of the Linguistic Society of America (LSA), Chicago, Illinois*.

Yuan, Q., Sclaroff, S., and Athitsos, V. (2005). Automatic 2D Hand Tracking in Video Sequences. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, volume 1, pages 250–256.

9. Language Resource References

Fenlon, J., Cormier, K., Rentelis, R., Schembri, A., Rowley, K., Adam, R., and Woll, B. (2014). *BSL SignBank: A lexical database of British Sign Language (First Edition)*. Deafness, Cognition and Language Research Centre, University College London. PID https://bslsignbank.ucl.ac.uk/.

Hilzensauer, M. and Krammer, K. (2015). *A multilingual dictionary for sign languages: "Spreadthesign"*. IATED. PID https://www.spreadthesign.com.

Hochgesang, J. A., Crasborn, O., and Lillo-Martin, D. (2023). *ASL Signbank*. PID https://aslsignbank.haskins.yale.edu/.

Kopf, M., Schulder, M., and Hanke, T. (2022). *The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources*. European Language Resources Association (ELRA). PID https://www.sign-lang.uni-hamburg.de/lrec/pub/22025.pdf.