# How Fast Is Sign Language? A Reevaluation of the Kinematic Bandwidth Using Motion Capture

Félix Bigand, E. Prigent, B. Berret, Annelies Braffort

# How Fast Is Sign Language? A Reevaluation of the Kinematic Bandwidth Using Motion Capture

Félix Bigand*, Elise Prigent*, Bastien Berret† and Annelies Braffort*
*Université Paris-Saclay*
**CNRS, LISN*
†*CIAMS, Institut Universitaire de France*
91400, Orsay, France
{felix.bigand, elise.prigent, bastien.berret, annelies.braffort}@universite-paris-saclay.fr

***Abstract*—Human motion lies within a range of low frequencies. Filtered and down-sampled motion capture (mocap) data can thus provide meaningful representations for computational models. However, little is known about the kinematic bandwidth of Sign Language (SL), apart from isolated signs. Studies examining isolated signs suggested that SL could be limited to relatively low frequencies. This is unlikely to be appropriate for real-life conditions where signs are produced faster and are combined with several other rapid motion features. The present study investigated the spectral content of a multi-signer mocap dataset of continuous signing in French Sign Language. Across six different signers, Power Spectral Density estimation and residual analysis of the mocap data revealed that SL motion can be limited to a 0–12-Hz bandwidth, which is substantially wider than state-of-the-art estimates on isolated signs. More specifically, filtering the movements below 6 Hz caused distortion of the rapid motion, which suggests that SL motion involves higher frequencies in real-life conditions. The importance of kinematic bandwidth estimation is further addressed with a machine learning model trained to identify the six signers of the dataset. The performance of the model significantly decreased when using inappropriate bandwidths.**

***Index Terms*—motion capture, motion analysis, spectral analysis, sign language**

## I. Introduction

Sign languages are used by 70 million deaf people worldwide and are specific to every country (over 200 different sign languages) [1]. Despite that, the majority of communication tools only rely on spoken or written languages. This raises an issue of accessibility as sign languages, unlike spoken ones, have no written form. For deaf persons, reading written content means reading a second language, which is not always mastered[1]. Many technological barriers must thus be tackled in order to provide accessible content in Sign Language (SL) for deaf users. This means developing models of automatic SL processing such as automatic SL translation [3], or SL avatars animation [4] [5] [6].

In order to design relevant models of SL, it is necessary to properly estimate the frequency content of SL movements, provided by motion capture (mocap). Mocap systems allow for the recording of a person's movements at high frame rates (e.g. 120, 250 frames per second (fps)). Then, mocap data are often filtered for human motion analyses [7] [8], which requires estimating an optimal cutoff frequency. Up to now, it is unclear what actual bandwidth should be taken to properly model SL motion. SL movements differ from non-linguistic ones, as they are constrained by not only biomechanic but also linguistic rules. More specifically for technological application perspectives, this problem must be answered in order to better understand whether the spectral content of SL motion is entirely represented when extracted from videos at low frame rates (e.g. 24 fps) [9].

The estimated bandwidth of human arm and head motion lies between 2 and 20 Hz, according to Bishop et al. [10]. More recently, Skogstad et al. also showed that rapid arbitrary motion of the hand may have an upper-bound frequency between 15 and 20 Hz [11]. The spectrum of SL motion has been investigated with isolated signs. Individual lexical signs were produced by one signer and taken out of context. Poizner et al. suggested that most of the energy of SL motion may lie below 6 or 7 Hz [12]. According to Foulds, a 0–3-Hz range is enough to understand American SL (ASL) isolated signs and fingerspelling (i.e. spelling out isolated words by producing letters with the hands) [13]. Sperling et al. also reported no significant intelligibility loss for ASL isolated signs from 30 to 10 fps, suggesting a 0–5-Hz bandwidth [14].

The limitation of these studies is that SL cannot be restricted to isolated signs taken out of context. Because of coarticulation, the duration of signs is shorter when produced in context rather than isolated [15] [16]. In addition to lexical signs, SL production is a continuous stream that involves multiple features, including rapid manual (e.g. pointing) and non-manual (e.g. eye gaze) movements. It can therefore be hypothesized that the actual bandwidth of SL motion is wider than previous estimates. As a matter of fact, one of the few studies that investigated real conversation conditions precisely indicated that 5 fps was too low for a comfortable SL conversation [17]. Additionally, the mentioned studies assessed the signed movements of one individual, which does not account for differences in speed between signers.

The aim of the present study was to overcome these limitations by evaluating the spectral content of continuous signing and over multiple signers (Section II). To this end, a two-step computational analysis of motion capture data was conducted

[1]"American deaf students around age 18 have a reading level more typical of 10-year-old hearing pupils" [2]

(Section III). Power Spectral Density estimation and residual analysis were applied to a mocap dataset of six signers in continuous French Sign Language. To the authors' knowledge, the present study is the first to use this computational workflow for SL. Moreover, the high precision of our mocap system allowed for the evaluation of higher frequencies compared with state-of-the-art studies (250 fps vs. 30 fps [13] [14]). In order to test the effects of the kinematic bandwidth estimation, bandwidth limited mocap data were used to train a machine learning model for person identification (Section IV). The performance of the model was assessed, as a function of the used bandwidth.

## II. The Motion Dataset

The data used for analysis was taken from a previously reported study [18] [19]. In brief, six signers each had freely described the content of 24 different pictures in French Sign Language (LSF). Based on a motion capture system (Optitrack S250e), the data consisted of the upper-body movements (19 markers) recorded at 250 fps, in three dimensions (Figure 1). From the 24 mocap recordings, only 21 were taken into account for the frequency content estimation (Section III), as three of them were already low-pass filtered for one signer. Each recording was a truncated version of the original one (5 sec), so that all the analyzed examples had the same length. Postures were normalized in order to filter out anthropometric differences, by substracting each signer's average posture from each frame, then adding the average posture over all signers.

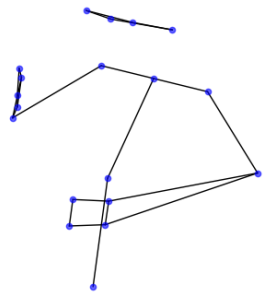

Fig. 1. The 19 upper-body markers recorded in the LSF mocap dataset.

## III. Frequency Content Estimation

### A. Power Spectral Density estimation

Similarly to Skogstad et al. [11], a two-step analysis was conducted in order to choose the optimal kinematic bandwidth. First, the frequency content of each signer's movements was estimated by measuring Power Spectral Density (PSD) using the Welch method [20]. Trajectories were split into overlapping segments over time, then the periodogram (i.e. magnitude squared of the windowed Discrete Fourier Transform) of each segment was computed. The PSD estimates were finally obtained by averaging the periodogram values over all segments. The present analysis was carried out using a Hann window of size 250 (1 sec), with 66% overlap.

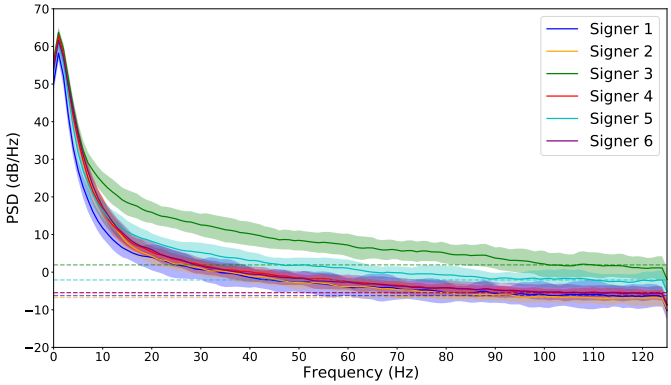


Fig. 2. Power Spectral Density estimation of the mocap data of 6 signers. Dashed horizontal lines indicate the noise floor estimate, for each signer. Error shaded regions indicate standard deviations over mocap examples.

The PSD estimates of each signer's mocap data are shown in Figure 2. The PSD values were averaged over body markers and over mocap examples. Noise floors were estimated by a visual analysis of the PSDs. Signers 1, 2, 4 and 6 reported similar noise floor (-6 dB/Hz), while it was higher for signer 5 (-2 dB/Hz) and additionally higher for signer 3 (+2 dB/Hz). Most of the power distribution lies between 0 and 5 Hz, with a 3-Hz peak, for all signers. Still, higher frequencies seem to contribute significantly as the associated PSD values are distinct from the noise floor up to 50 Hz (or higher).

### B. Residual analysis

To further understand whether these higher frequencies related to actual motion information or measurement noise, a residual analysis was conducted [21]. This method consists of measuring the average difference between the unfiltered and filtered signal, over several cutoff frequencies. In this study, the mocap data were low-pass filtered using a fourth-order Butterworth filter.

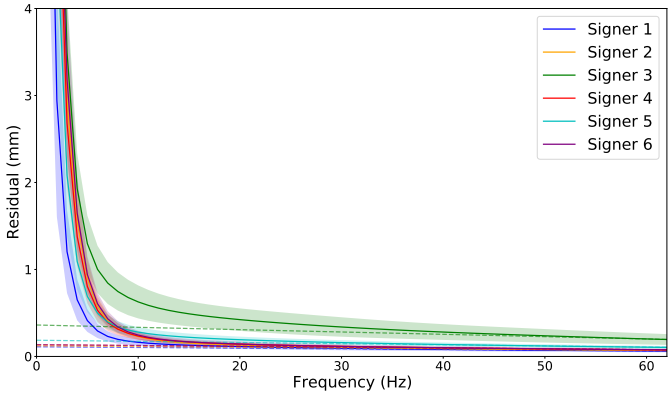


Fig. 3. Residual plot between the unfiltered and filtered mocap data of 6 signers, as a function of the filter cutoff frequency. Dashed lines indicate the noise residual estimate, for each signer. Error shaded regions indicate standard deviations over mocap examples.

Results of the residual analysis between unfiltered and filtered mocap data are displayed in Figure 3. As for PSD

estimation, the residual values were averaged over markers and examples. The estimates of noise residuals were obtained by defining the regression line from 40 Hz to $fps/2$ Hz. Indeed, theoretically, the residual curve of noise is a linear curve from an intercept at 0 Hz to the ($fps/2$ Hz, 0 mm) point.

This 0-Hz intercept provides an estimate for the Root Mean Square (RMS) of noise, following the definition of Winter [21]. In other words, this value reflects the mean displacement of sensors caused by measurement noise. Estimated RMS of noise were similar for signer 1 (0.11 mm), signer 2 (0.12 mm), signer 4 (0.14 mm) and signer 6 (0.14 mm). Higher values were reported for signer 5 (0.19 mm) and additionally higher for signer 3 (0.37 mm). These results confirmed the PSD estimation, suggesting that the mocap data of signer 3 and 5 were the noisiest.

Interestingly, Figure 3 shows that the residual values relating to signer 5 (cyan curve) are lower compared with signers 2, 4 and 6 (yellow, red and magenta curves) in low frequencies (below 10 Hz), but higher in high frequencies (above 10 Hz). The latter high-frequency comparison is in line with the noise RMS calculations. This suggests that most of the actual motion information of signer 5 may be in lower frequencies (i.e. slower movements), while his mocap recording is noisier.

### C. Choosing an optimal bandwidth

We then assessed different cutoff frequencies in order to define the optimal kinematic bandwidth of our data. Based on prior work, three cutoff frequencies were compared: 6, 12 and 25 Hz. The lower frequency relates to a 1-mm residual[2], which was reported as being an imperceptible deviation for arbitrary hand motion [11]. The upper one is the frequency for which the residual equals the noise RMS[2]. Using this cutoff value, the signal distortion should equal the amount of noise allowed through [21]. Finally, 12 Hz is an intermediate value which is of great interest as it would be the highest cutoff frequency possible with most video systems (using 24 fps), following the Nyquist-Shannon theorem [22] [23].

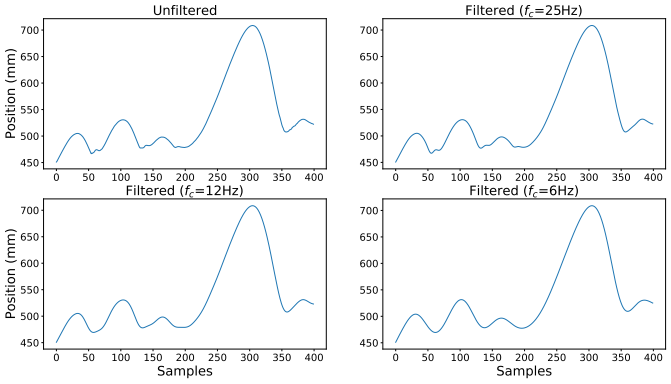


Fig. 4. Example of slow motion: Z-axis trajectory of the right hand of Signer 5, example 5. Subplots allow for comparison between unfiltered and filtered ($f_c$ = 25, 12 or 6 Hz) mocap data.

[2]When different frequencies were possible across individuals, the maximum frequency was chosen, in order to minimize signal distortion.

When looking at slow motion (Figure 4), the 6-Hz and 12-Hz filters seem to be optimal solutions for denoising the data. The 6-Hz filter might even be slightly more promising, as it filters out more artifacts than the 12-Hz one. However, this finding is not confirmed with rapid motion (Figure 5), where filtering at 6 Hz cancels important fast movements. For instance, the residual values relating to signer 3 almost double from 6 Hz (0.56 mm) to 12 Hz (1.00 mm). Interestingly, the 12-Hz filter smoothes out most of the signal, while keeping fast oscillations intact. Filtering at 25 Hz instead of 12 Hz does not seem to add substantial information and differences in the residuals are negligible (mean = 0.08 mm, std = 0.04).

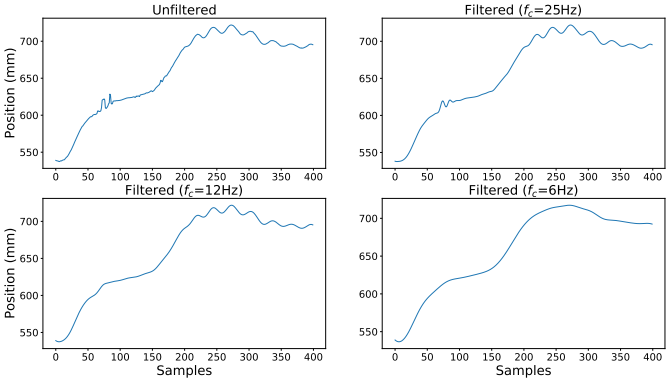


Fig. 5. Example of rapid motion: Z-axis trajectory of the right hand of Signer 3, example 17. Subplots allow for comparison between unfiltered and filtered ($f_c$ = 25, 12 or 6 Hz) mocap data.

According to these results, the conclusions about the 3 proposed bandwidths for SL motion are summarized as follows:

- 0–6 Hz : The main movements are captured and the noise artifacts are drastically reduced. However, rapid motion is filtered out, which distorts the motion representation for faster signers.
- 0–12 Hz : The fastest movements are captured while the noise artifacts are still importantly reduced.
- 0–25 Hz : Noise artifacts are allowed through, while the additional information compared with a 0–12-Hz range is negligible, as regards residual results.

### D. Discussion

Based on the combined results of residual analysis and data visualization, 0–12 Hz was found to be a reasonable bandwidth for our SL mocap dataset. This is noticeably wider than the 0–3-Hz [13] or 0–6-Hz [12] previously reported bandwidths. More specifically, a 0–6-Hz range was not able to account for the rapid signed motion. This range was associated with a 1-mm residual, which was reported to be an unsignificant distortion for rapid arbitrary motion [11]. A 1-mm deviation may thus not be negligible for SL motion, suggesting that this latter contains finer movements. This is consistent as, compared to arbitrary hand motion, SL obeys to specific linguistic constraints and requires precise movements of the hands and fingers for comprehensibility [24]. Moreover, the

precision of the analyzed motion may also have been caused by the high level of expertise of the six signers.

It was not clear whether a 0–25-Hz bandwidth actually provided additional information about the real motion rather than noise. Although it was negligible here, it might be possible that higher frequencies relate to actual motion, particularly for fast signers. Further work measuring eye movements could also refine the relevance of a wider bandwidth. To emphasize the need for an optimal cutoff frequency, we assessed the performance of a machine learning model, when trained on filtered or unfiltered inputs.

## IV. Effects of the Kinematic Bandwidth Estimation

### A. Feature extraction: velocity and acceleration

Computational models rely on features extracted from the trajectories of markers, such as velocity or acceleration. Figure 6 illustrates the advantage of applying a reasonable low-pass filtering. At each step of differentiation, the amount of noise is amplified. More interestingly, without filtering, the acceleration data are almost not readable, which may cause wrong interpretations of inter-individual differences (e.g. acceleration peak of signer 5, instead of signer 3). Person identification is particularly suited to further illustrate this issue, as wrong interpretations of inter-individual differences may cause wrong predictions of the identified person. Moreover, person identification from motion recently raised important social issues about the confidentiality of individuals in Sign Language [18]. Therefore, using person identification as an example, a machine learning model was trained and its performance was assessed, as a function of the used bandwidth.

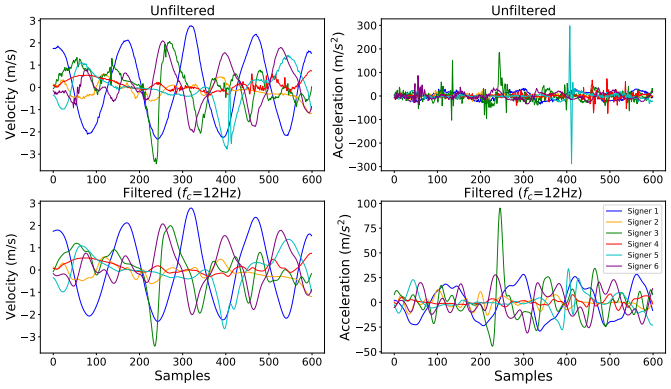


Fig. 6. Z-axis velocity (left) and acceleration (right) curves of the right hand, for all signers, example 1. Subplots allow for comparison between unfiltered and 12-Hz filtered mocap data. Note the scale difference of the Y dimension for acceleration, revealing the unrealistic values caused by unfiltered data.

### B. Machine learning model

The model was designed using a statistical-based approach. Statistics (mean and standard deviation) were measured from temporal features of each mocap example (Section II). The used temporal features were a combination of local position, velocity and acceleration. The identification step consisted of applying Principal Component Analysis to the statistics $(\mu_{pos}, \sigma_{pos}, \mu_{vel}, \sigma_{vel}, \mu_{acc}, \sigma_{acc})$ of all examples and finally training a multinomial logistic regression model on the extracted Principal Components (PCs). The model was trained iteratively on N-1 (23) examples for each signer, and the remaining 1 observation was used as test exemplar. Performance was computed as an average over the 24 test iterations.

To illustrate the impact of the bandwidth estimation on our model, the first extracted PC was analyzed. This component was highly correlated with global dynamic statistics $(\sigma_{vel}, \sigma_{acc})$ for both unfiltered $(r(16416) = .63, p < .001)$ and 12-Hz filtered $(r(16416) = .65, p < .001)$ inputs[3]. However, the model provided different results depending on the filtering step. Figure 7 displays the first 2 PCs scores and the model confusions when trained on PC1.

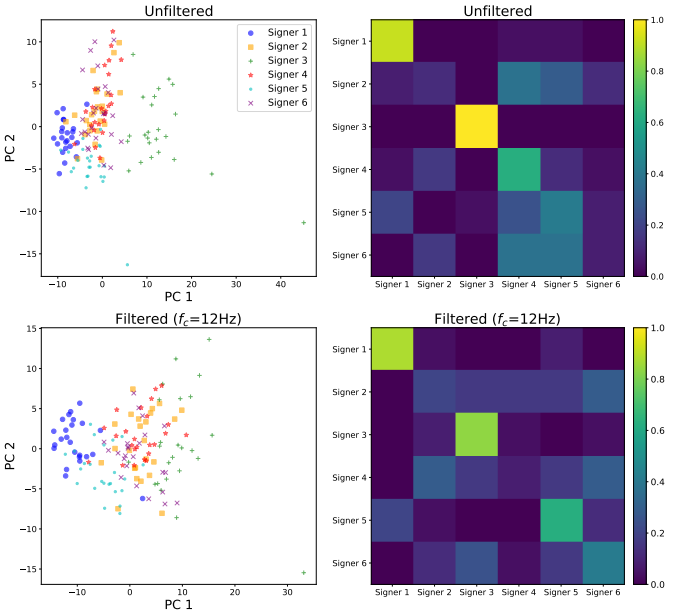


Fig. 7. Projection of all the mocap examples over the first 2 PCs extracted by the model (left). Confusion matrix of identifications (averaged over 24 tests), when the model is trained only on PC1 (right). The two rows allow for comparison between unfiltered and 12-Hz filtered mocap data.

For unfiltered mocap data, the PC1 scores of signer 5 (mean = -0.15) were confused with signer 2 (mean = -1.43), signer 4 (mean = -2.72) and signer 6 (mean = -1.04). By contrast, signer 1 had lower scores (mean = -7.72) and signer 3 was surprisingly highly separated from the 5 other signers (mean = 13.07). The performance of the model trained on PC1 confirmed this idea, as signer 1 (91.7 %) and signer 3 (100%) were correctly identified, by contrast with signer 5 (41.7 %). However, when applying a 12-Hz filtering, the PC1 scores were different. Signer 1 still had the lowest scores (mean = -10.4), but signer 5 (mean = -5.05) was separated from signer 2 (mean= 2.17), signer 4 (mean = 1.99) and signer 6 (mean = 1.67). The highest scores were still reported for signer 3 (mean

[3]By contrast, the correlation with static statistics $(\mu_{pos})$ was not significant for unfiltered $(r(8208) = -.02, p = .14)$ and 12-Hz filtered $(r(8208) = -.02, p = .18)$ inputs.

= 9.62), but the gap with other signers was lower. This time, the model succeeded in identifying signer 5 with a higher score (62.5%) based on his dynamic differences, and signer 3 was still identified (83.7%) but potentially based on a more realistic interpretation of PC1. Finally, when the model was trained on 6-Hz filtered motion, the identification accuracy significantly decreased for the fastest signer, signer 3 (70.8 %).

### C. Discussion

These results reflect the observations made in Section III. With the wrong bandwidth estimation, our model may have misidentified signer 5 because of slower but also noisier motion data, compared to signers 2, 4 and 6. Similarly, the separation of signer 3 from other signers was surprisingly wide, which may have been caused by the fact that the mocap of signer 3 contained the fastest but also noisiest data. The results using a 6-Hz filter also confirm that a 0–6-Hz range distorts rapid motion and thus provides an incomplete representation for models. Based on the example of PC1, 12-Hz filtered data provided the best representation to correctly differentiate between the dynamics of the six signers.

## V. Conclusion and discussion

The present study provides a new estimate of the kinematic bandwidth of Sign Language using computational methods, as was done for gait and hand motion [11] [21]. Compared to prior work on SL [12] [13] [14], our results suggest that SL motion contain higher frequencies (0–12 Hz). In prior studies, only isolated signs or fingerspelling had been investigated. In the present study, signers had freely described pictures in French Sign Language without any constraints in time, signs or structure. These results thus support the hypothesis that signing may be faster when it is done in context, rather than when it is isolated [15]. Additionally, the use of six signers' mocap allowed for more generalization, compared to prior work. Finally, the analysis of a machine learning model emphasizes the need for a correct filtering of mocap data when designing SL models [25] [5] [3]. For technological application purposes, these results support the potential for SL motion data extracted from videos at 24 fps [9]. Despite the fact that estimating movements from videos remains limited to two dimensions, it may properly capture spectral information, following Nyquist-Shannon rule. These outcomes call for additional research further investigating the kinematic bandwidth of SL, across other signers and within different linguistic contexts.

## Acknowledgment

This work has been funded by the Bpifrance investment project "Grands défis du numérique", as part of the ROSETTA project (RObot for Subtitling and intElligent adapTed TranslAtion).

## References

[1] W. F. of the Deaf. (2016) Our work. [Online]. Available: http://wfdeaf.org/our-work/

[2] J. A. Holt, "Stanford achievement test—8th edition: Reading comprehension subgroup results," *American Annals of the Deaf*, vol. 138, no. 2, pp. 172–175, 1993.

[3] V. Belissen, "From sign recognition to automatic sign language understanding: Addressing the non-conventionalized units," Ph.D. dissertation, Université Paris-Saclay, 2020.

[4] M. Filhol, J. Mcdonald, and R. Wolfe, "Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2017, pp. 27–40.

[5] S. Gibet, "Building french sign language motion capture corpora for signing avatars," 2018.

[6] L. Naert, C. Larboulette, and S. Gibet, "Lsf-animal: A motion capture corpus in french sign language designed for the animation of signing avatars," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 6008–6017.

[7] M. Zago, M. Codari, F. M. Iaia, and C. Sforza, "Multi-segmental movements as a function of experience in karate," *Journal of Sports Sciences*, vol. 35, no. 15, pp. 1515–1522, 2017.

[8] E. Carlson, P. Saari, B. Burger, and P. Toiviainen, "Dance to your own drum: Identification of musical genre and individual dancer from motion capture using machine learning," *Journal of New Music Research*, pp. 1–16, 2020.

[9] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[10] G. Bishop, G. Welch, and B. D. Allen, "Tracking: Beyond 15 minutes of thought," *SIGGRAPH Course Pack*, vol. 11, 2001.

[11] S. A. v. D. Skogstad, K. Nymoen, M. E. Høvin, S. Holm, and A. R. Jensenius, "Filtering motion capture data for real-time applications," 2013.

[12] H. Poizner, E. S. Klima, U. Bellugi, and R. B. Livingston, "Motion analysis of grammatical processes in a visual-gestural language," *Event cognition: An ecological perspective*, pp. 155–174, 1986.

[13] R. A. Foulds, "Biomechanical and perceptual constraints on the bandwidth requirements of sign language," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 1, pp. 65–72, 2004.

[14] G. Sperling, M. Landy, Y. Cohen, and M. Pavel, "Intelligible encoding of asl image sequences at extremely low information rates," *Computer vision, graphics, and image processing*, vol. 31, no. 3, pp. 335–391, 1985.

[15] A. Braffort, L. Bolot, and J. Segouat, "Virtual signer coarticulation in octopus, a sign language generation platform," in *GW 2011: The 9th International Gesture Workshop*, 2011.

[16] C. C. Koech, "A kinematic analysis of sign language," 2006.

[17] N. Cherniavsky, A. C. Cavender, R. E. Ladner, and E. A. Riskin, "Variable frame rate for low power mobile sign language communication," in *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, 2007, pp. 163–170.

[18] F. Bigand, E. Prigent, and A. Braffort, "Person identification based on sign language motion: Insights from human perception and computational modeling," in *Proceedings of the 7th International Conference on Movement and Computing*, 2020, pp. 1–7.

[19] M.-e.-F. Benchiheub, B. Berret, and A. Braffort, "Collecting and Analysing a Motion-Capture Corpus of French Sign Language," in *Workshop on the Representation and Processing of Sign Languages*, Portoroz, Slovenia, Jan. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01633625

[20] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[21] D. A. Winter, *Biomechanics and motor control of human movement*. John Wiley & Sons, 2009.

[22] H. Nyquist, "Certain topics in telegraph transmission theory," *Transactions of the American Institute of Electrical Engineers*, vol. 47, no. 2, pp. 617–644, 1928.

[23] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[24] H. Poizner, U. Bellugi, and V. Lutes-Driscoll, "Perception of american sign language in dynamic point-light displays." *Journal of experimental psychology: Human perception and performance*, vol. 7, no. 2, p. 430, 1981.

[25] E. Malaia, R. B. Wilbur, and M. Milković, "Kinematic parameters of signed verbs," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1677–1688, 2013.