Synthesis for the Kinematic Control of Identity in Sign Language

Félix Bigand, Elise Prigent, Annelies Braffort

Synthesis for the Kinematic Control of Identity in Sign Language

Félix Bigand, Elise Prigent, Annelies Braffort
Université Paris-Saclay, CNRS, LISN, Orsay, France
{felix.bigand, elise.prigent, annelies.braffort}@lisn.upsaclay.fr

Abstract

Sign Language (SL) animations generated from motion capture (mocap) of real signers convey critical information about their identity. It has been suggested that this information is mostly carried by statistics of the movements kinematics. Manipulating these statistics in the generation of SL movements could allow controlling the identity of the signer, notably to preserve anonymity. This paper tests this hypothesis by presenting a novel synthesis algorithm that manipulates the identity-specific statistics of mocap recordings. The algorithm produced convincing new versions of French Sign Language discourses, which accurately modulated the identity prediction of a machine learning model. These results open up promising perspectives toward the automatic control of identity in the motion animation of virtual signers.

Keywords: Sign Language, Anonymized Content, Identity Conversion, Motion Generation, Machine Learning

1. Introduction

Using motion capture (mocap) systems, the movements of signers can be recorded with high accuracy and be used to produce natural and comprehensible content (Lu and Huenerfauth, 2010; Gibet, 2018). However, this process raises an unexpected problem, related to the human ability to identify individuals from their movements (Troje et al., 2005; Loula et al., 2005; Bläsing and Sauzet, 2018). As for spoken languages in the auditory domain, where voice parameters inform about the identity of a speaker, signers can be identified from their movements (Bigand et al., 2020). We present a synthesis algorithm for controlling the motion features that characterize the identity of a signer. This would allow producing anonymized, non-identifiable, content with virtual signers, which is crucial (e.g., for sharing anonymized testimony) given that Sign Languages (SLs) have no written form (Lee et al., 2021).

In line with prior work on non-SL movements (Troje et al., 2005; Carlson et al., 2020; Zhang and Troje, 2005), our recent studies suggested that identity was mainly inferred from the kinematic aspects of the movements, beyond size, shape or posture of the signers (Bigand et al., 2020; Bigand et al., 2021). Using a machine learning model, we automatically extracted the specific kinematic aspects of motion that carry identity, using time-averaged statistics (Section 2). The present synthesis algorithm was then developed in order to manipulate the identity-specific statistics of original mocap recordings (Section 3). We tested the performance of the synthesis algorithm by modifying the identity attribute of mocap recordings in French Sign Language, and by assessing the identity inferred from the new excerpts (Section 4). This constitutes the first step toward automatically anonymizing the movements of signers in SL animations, in the same way as for the voice of a speaker, which can be anonymized by modifying specific vocal parameters (Section 5).

2. Motion statistics of identity

Mocap recordings were taken from MOCAP1 corpus (Benchiheub et al., 2020). Six signers had freely described the content of 24 pictures using French Sign Language (LSF). From each of the 24 original recordings, one mocap recording unit of 5-seconds duration was extracted from the beginning of the utterance (see examples in Videos 7.1 to 7.6). As shown in Figure 1, the used markers were (L = left, R = right, F = front, B = back): (1) pelvis, (2) stomach, (3) sternum, (4) LB head, (5) LF head, (6) RB head, (7) RF head, (8) L shoulder, (9) L elbow, (10) LB wrist, (11) LF wrist, (12) LB hand, (13) LF hand, (14) R shoulder, (15) R elbow, (16) RB wrist, (17) RF wrist, (18) RB hand, (19) RF hand. The mocap examples were normalized with respect to size, shape and posture of the signers (see Bigand et al. (2021)). The mocap data of the pelvis marker were ignored as it was set as the origin, which leads to zero vectors. Position and velocity of the body markers were used as temporal features. Velocity was estimated by time differentiation of the mocap position coordinates.

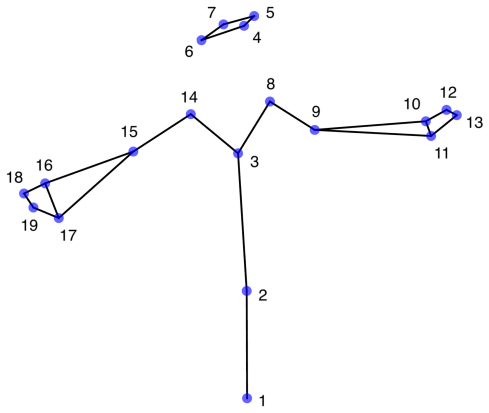

Figure 1: The 19 upper-body markers used in the mocap recordings.

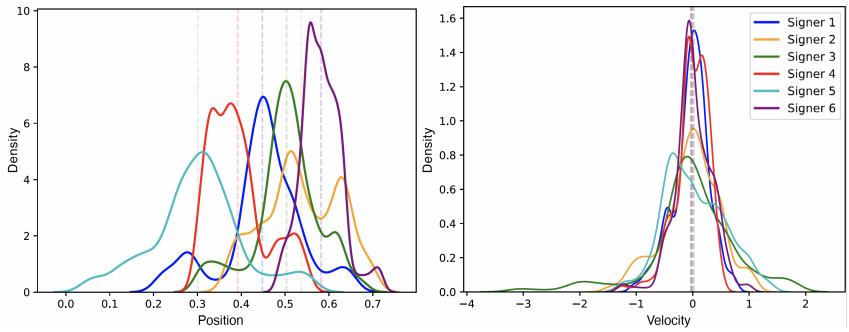

Figure 2: Distributions of position and velocity data of the RF hand marker along the Z axis, for mocap example 24. Dashed vertical lines represent the means.

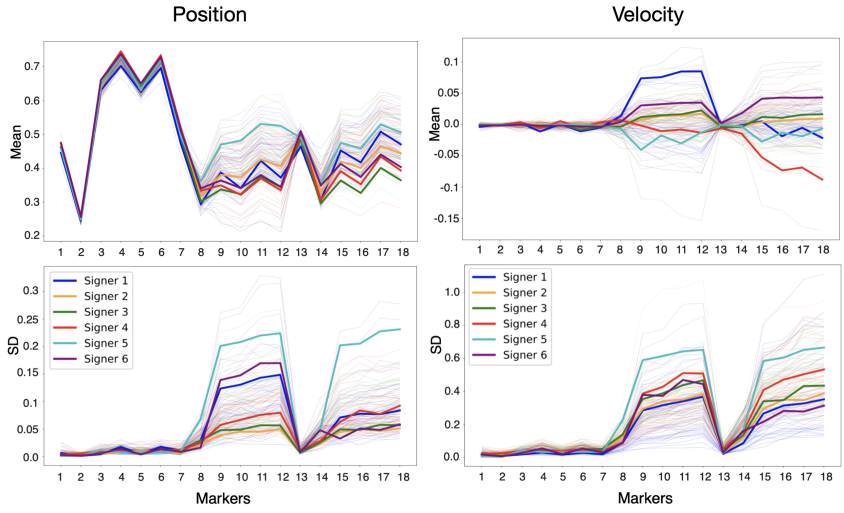

Figure 3: The two moments of the position and velocity data along the Z axis, for all markers and all 144 mocap examples. Thick lines represent the average statistics of each signer across their 24 examples.

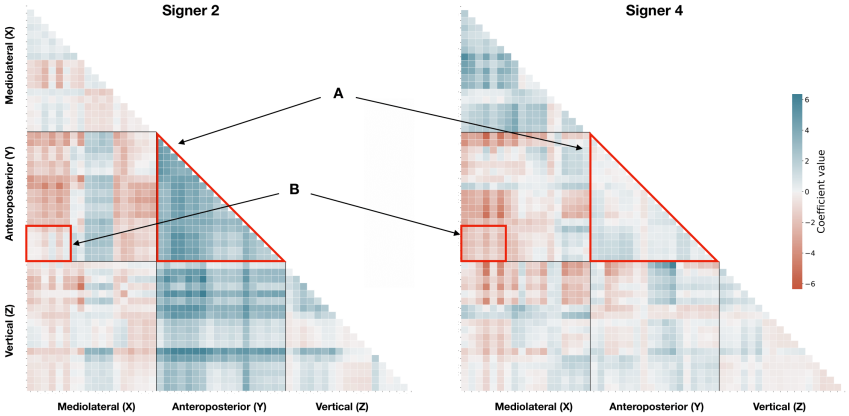

Figure 4: The covariance of velocity between body markers (rows and columns) of Signer 2 and Signer 4 in the three dimensions, for mocap example 24. Markers are sorted from the 1^{st} to the 19^{th} as presented in Bigand et al. (2021), along X, Y and Z axes. Coefficients correspond to the covariance measures centered and standardized across examples and signers. Blue represent positive covariances, while red represent negative ones. (A) covariance between all markers along the Y axis. (B) covariance between the right hand and arm markers along the Y axis, and the trunk and head markers along the X axis.

Statistics of the mocap examples were then computed as follows. Based on previous research investigating the perception of auditory and visual textures (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000), we measured the first two moments (i.e., mean and standard deviation (SD)) of position and velocity, and covariances of velocity between body markers. The first two moments of position and velocity described their statistical distributions, which may vary from one individual to another, as shown for expert gesture analysis (Tits, 2018). Moreover, the covariance of velocity allowed for quantifying the extent to which any two markers covaried with each other, in two directions. This latter statistic has been shown to allow for automatic person identification from dance movements (Carlson et al., 2020).

These statistics vary substantially across the mocap data of different signers. For instance, as shown in Figure 2, the position and velocity data of one body marker are distributed differently across signers, for one mocap example (i.e., for comparable content: the description of the same picture in LSF). Distributions of position data differ in location of the peak (captured by the mean) and width (captured by the standard deviation). Figure 3 further supports that the two moments of position and velocity may capture substantial differences across signers.

Furthermore, velocity covariances capture different aspects of motor coordination between the markers in three dimensions, which can differ across signers. Various distinct coordination patterns can be extracted. For instance, for mocap example 24, the movements of Signer 2 show an overall substantial (positive) covariance between body markers along the Y axis, while this covariance is near zero for Signer 4 (Figure 4.A). Inversely, Signer 4 displays an important (negative) covariance of movements of the right arm and hand along the Y axis with the trunk (i.e., stomach and sternum) and head markers along the X axis, while this covariance is less important for Signer 2 (Figure 4.B). Taken together, these examples raise the possibility that the identity of a signer is conveyed by statistical properties of his or her movements.

3. Methods

The automatic signer identification model presented in Bigand et al. (2021) allowed extracting specific kinematic statistics that carry identity information about the signers. A linear classifier was trained to extract the statistics of the mocap data characteristic of identity (i.e., the ones that allow for accurate signer identification).

Then, the aim of the present synthesis algorithm was to manipulate the statistics of an original SL mocap recording (i.e., impose new statistics to the original recording), in order to reduce ($\alpha < 0$) or exaggerate ($\alpha > 0$) the identity attribute, following Equation 1:

$$\tilde{\mathbf{d}}_\alpha = \mathbf{d_{orig}} + \alpha \mathbf{d_k} \quad (1)$$

where $\tilde{\mathbf{d}}_\alpha$ is a vector containing the new target statistics to be imposed by the synthesis alogrithm, $\mathbf{d_{orig}}$ is a vector containing the original statistics of the mocap example, $\mathbf{d_k}$ is a vector containing the overall statistical patterns characteristic of the identity of Signer k, and α is a scalar related to the amount of reduction ($\alpha < 0$) or exaggeration ($\alpha > 0$) of the identity attribute.

The different steps of the synthesis process are displayed in Figure 5. In summary, the synthesis process consisted of modifying (i.e., "re-synthezing") an existing mocap recording in order to change the identity attribute of the signer, according to the following steps. First, statistics of the original mocap example are measured, while the discriminant statistical kinematic patterns are extracted by the automatic identification model (see Bigand et al. (2021)). Then, the discriminant statistics characteristic of Signer k are either added to ($\alpha > 0$) or subtracted from ($\alpha < 0$) the ones of the original example (see Equation 1). Multiple manipulations can then be done using this technique, depending on the values of k and α. For instance, if the original mocap example relates to Signer 1, reducing the importance of her identity-specific statistics (i.e., $k = 1, \alpha < 0$) would make her less identifiable (i.e., kinematic anonymization). By contrast, increasing the importance of the identity-specific statistics of Signer 2 (i.e., $k = 2, \alpha > 0$) would make this latter signer identifiable while the SL movements were originally executed by Signer 1 (i.e., kinematic identity conversion). Once the target statistics defined, they are imposed to the original mocap signal by the algorithm, which creates a new mocap excerpt.

Target statistics were imposed using an iterative process where a synthesized mocap signal (initialized with the content of the original mocap recording) is modified until its statistics are sufficiently close to the target ones $\tilde{\mathbf{d}}_\alpha$. Mathematically, the objective of this process is to minimize the loss function that calculates the mean square of the differences between the target statistics and the statistics of the synthesized movements (see Equation 2). We imposed the first two moments (mean and SD) of position and velocity data and the covariance of velocity between markers, as they were found to be the most important statistics for signer identification (Bigand, 2021). Imposing the mean of position and mean of velocity of the markers was done to maintain consistent motion data when synthesizing (e.g., to avoid the generation of unrealistic, non-biological, movements), although these two statistics had only minor role in the identification.

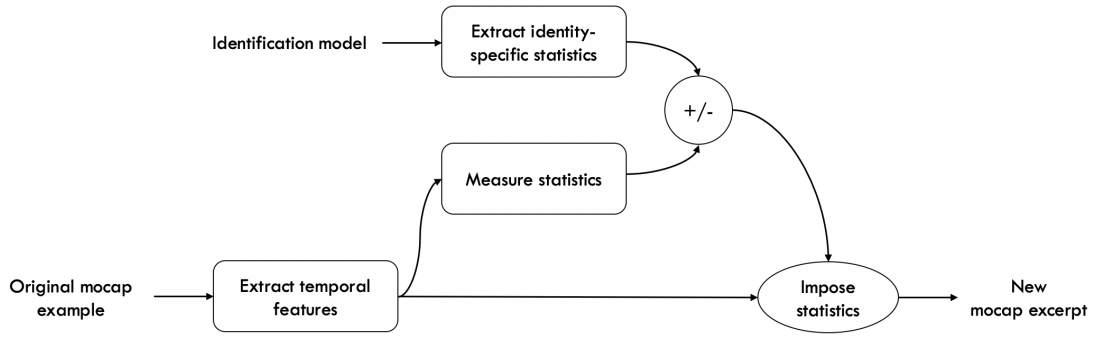

Figure 5: Schematic representation of the steps used in the synthesis algorithm for the kinematic control of identity.

$$
\begin{aligned}
loss_1 &= \sum_m (\mu_{pos,m,targ} - \mu_{pos,m,synth})^2 \\
loss_2 &= \sum_m (\sigma_{pos,m,targ} - \sigma_{pos,m,synth})^2 \\
loss_3 &= \sum_m (\mu_{vel,m,targ} - \mu_{vel,m,synth})^2 \\
loss_4 &= \sum_m (\sigma_{vel,m,targ} - \sigma_{vel,m,synth})^2 \\
loss_5 &= \sum_{i,j} (C_{i,j,targ} - C_{i,j,synth})^2 \\
loss_{tot} &= \sum_{i=1}^{5} loss_i
\end{aligned}
\tag{2}
$$

where $\mu_{pos,m}$, $\sigma_{pos,m}$, $\mu_{vel,m}$ and $\sigma_{vel,m}$ are the first two moments of position and velocity data of marker m ($m \in [1, 54]$), $C_{i,j}$ is the covariance of velocity between markers i and j. $targ$ and $synth$ subscripts distinguish between target statistics and statistics of the synthesized movements, respectively.

In order to be able to minimize all of the five loss components of Equation 2 despite the differences in ranges of amplitude across statistics, we used a weighted loss function, whose weights then need to be optimized (see Equation 3). The loss function was then minimized using the Adam optimization algorithm for gradient descent. Each iterative step of the gradient descent modified the synthesized mocap signals (i.e., position temporal curves of the 19 markers along the three dimensions) so that they approached the target statistics.

$$
loss_{tot} = \sum_{i=1}^{5} w_i loss_i \tag{3}
$$

where $loss_i$ is the loss function related to one statistical measure and w_i the optimized weight.

Initially, there was no constraint in the synthesis process that forced the position and velocity signals of the synthesized movements to remain consistent with their initial temporal structure in the original movements. The limitation of this first version of the algorithm is that, although it managed to impose the statistics present in Equation 2, the modifications applied to the new movements seemed to generate noise artifacts rather than changing relevant aspects of the motion of the signer (see Video 10.1). In fact, the imposing algorithm managed to impose the target statistics but by modifying the movements in an undesired manner. First, low-energy segments of the motion were modified in the same way as high-energy ones, which is not relevant as they may not be perceived by observers. Moreover, reaching the target statistics caused very rapid oscillations in the synthesized velocity temporal curves, which are unlikely to be perceived as biological motion by the observers (but rather noisy, wobbling, markers).

In order to modify the movements in proportion to their energy (i.e., modify the aspects of the movement at relevant times of actual, perceptible, motion), we included another target statistic in the imposing algorithm: the correlation of velocity between the original and synthesized movements. The algorithm then aimed to minimize the mean squared error between this correlation and a value of 1, which characterizes two signals that are perfectly positively correlated (see Equation 4). In other words, imposing this additional statistic (Equation 5) allowed forcing the velocity curves of the synthesized movements to be consistent with their initial temporal structure in the original mocap recording (Figure 6).

$$
loss_6 = \sum_m (1 - \rho_{vel,m,synth})^2 \tag{4}
$$

$$
loss_{tot} = \sum_{i=1}^{6} w_i loss_i \tag{5}
$$

where $\rho_{vel,m}$ is the correlation of velocity between the original and synthesized movements of marker m ($m \in [1, 54]$). The target correlation value is set to 1 for all markers, in order to preserve the original temporal structure of velocity curves.

4. Results

This synthesis procedure was run on mocap examples of different signers and for different modifications of the identity attribute. In order to visualize how these

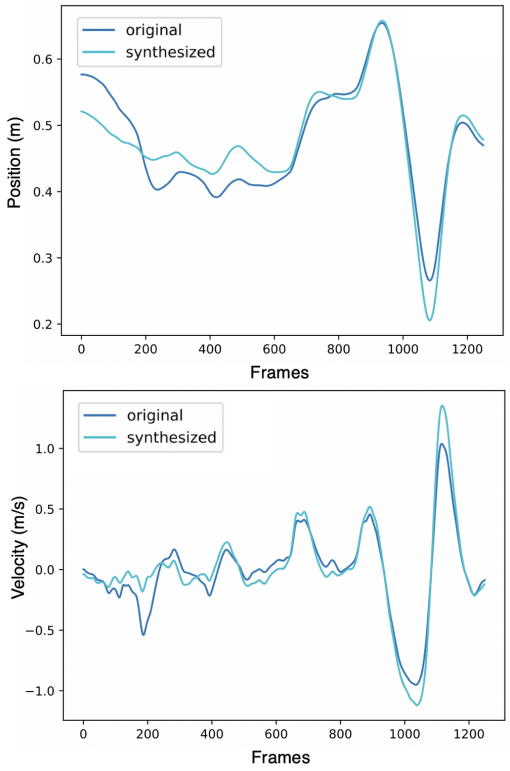

Figure 6: Example of the synthesis results for the identity conversion from Signer 1 to Signer 2 (mocap example 1). Position (up) and velocity (down) data of RF hand marker along the Z axis are shown, for the original mocap recording and synthesized mocap excerpt.

new statistics affected the movements of the SL discourse of Signer 1, the original and synthesized mocap examples can be seen as "point-light" display videos. For instance, the movements of Signer 1 were modified so that the perceived identity was that of Signer 2 (i.e., identity conversion) (see Videos 10.3 and 10.4). Then, they were modified to make Signer 1 not identifiable, without making another signer identifiable specifically (i.e., anonymization) (see Videos 10.5 and 10.6). One further synthesis example of identity conversion (from Signer 2 to Signer 1) can be found in Bigand (2021).

In order to assess the extent to which the novel movements generated by our algorithm could convey a modified identity attribute (e.g., could be anonymized, or identified as movements of another signer), we tested our automatic signer identification model on the synthesized mocap examples. If the identity-specific aspects of the movements are correctly modified by the synthesis algorithm, then automatic identification from these synthesized examples should be compromised.

When converting the identity of Signer 1 into that of Signer 2, the automatic signer identification model identified the synthesized mocap example as that of Signer 2, while it identified the original motion as produced by Signer 1 (see Table 1). Then, when anonymizing the content of Signer 1, the signer identification model did not manage to identify Signer 1 from the synthesized movements (see Table 1). Moreover, the highest identification probability from this excerpt was 0.43, which means that it did not clearly identify any other signer from the anonymized movements.

Table 1: Output of the automatic signer identification model from original and synthesized movements of Signer 1. The synthesized versions consist of identity conversion into Signer 2 and anonymization. Each output number is the probability that the movements were produced by the signer. Bold numbers represent the highest probability across the six signers.

	Original	Synthesized	
		Conversion	Anonymization
Signer 1	**0.99**	0.00	0.05
Signer 2	0.00	**0.99**	0.34
Signer 3	0.00	0.00	0.14
Signer 4	0.00	0.00	0.01
Signer 5	0.00	0.00	0.02
Signer 6	0.00	0.00	**0.43**

5. Discussion

This paper shows that simple statistics of the movements of a signer can be manipulated in order to regenerate mocap recordings with a modified identity attribute. The mocap data of SL discourses can undergo various manipulations, such as kinematic identity conversion or anonymization. Moreover, the synthesis algorithm preserves the original temporal structure of the movements, which is crucial because degrading temporal structure could impair the comprehension of the SL discourse.

Up to now, anonymization methods of SL content were modifying appearance, using virtual signers (Kipp et al., 2011) or modified videos (e.g., face-swapped videos, where the face of the signer is replaced with another face) (Lee et al., 2021; Bragg et al., 2020). Our technique focuses on controlling the identity in the kinematics of the signers, which could interestingly complement prior approaches in order to provide full anonymity, beyond face or body shape manipulations. Moreover, the proposed algorithm has the advantage that it can render the movements of signers as neutral (i.e., not reflecting the identity of any other signer), by contrast with face-swapping techniques.

However, some limitations of the present work should be noted in order to ensure an effective use of these tools in actual applications. First, although we aimed to use SL mocap data as representative as possible of real-life conditions (i.e., spontaneous LSF), the discourses used in the present study were picture descriptions, which may have involved specific linguistic structures more than others (e.g., depicting ones). The different outcomes reported here should be further tested in a wider linguistic context.

Moreover, the present computational findings call for further tests with human participants. Three key problems should be investigated, similarly to prior work on video anonymization (Lee et al., 2021): (1) identifiability, by verifying that the ability of human observers to identify the signers is compromised when showing the synthesized modified movements, as compared to the original ones (e.g., with "point-light" displays like in Bigand et al. (2020) and Troje et al. (2005)); (2) comprehensibility, by evaluating the extent to which the observers still understand the SL content in the modified motion examples; and (3) acceptability, by assessing the deaf user perspective on the virtual signers animated with the modified movements and discussing potential use cases (e.g., with focus groups). Should these three fundamental points be validated, the present work could constitute a first step of interest toward automatically controlling the identity of deaf SL users when expressing themselves via virtual signers. Moreover, as shown for videos (Bragg et al., 2020), preserving anonymity in mocap recordings could increase willingness of SL users to participate in mocap research (e.g., in data collection), which is crucial to develop effective and acceptable technologies.

6. Acknowledgements

This work has been funded by the Bpifrance (`https://www.bpifrance.fr/`) investment project "Grands défis du numerique", as part of the ROSETTA project (RObot for Subtitling and intElligent adapTed TranslAtion).

7. Bibliographical References

Bigand, F., Prigent, E., and Braffort, A. (2020). 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*, pages 1–7.

Bigand, F., Prigent, E., Berret, B., and Braffort, A. (2021). Machine learning of motion statistics reveals the kinematic signature of a person's identity in sign language. *Frontiers in Bioengineering and Biotechnology*, 9:603.

Bigand, F. (2021). *Extracting human characteristics from motion using machine learning : the case of identity in Sign Language*. Theses, Université Paris-Saclay, November.

Bläsing, B. E. and Sauzet, O. (2018). My action, my self: Recognition of self-created but visually unfamiliar dance-like actions from point-light displays. *Frontiers in psychology*, 9:1909.

Bragg, D., Koller, O., Caselli, N., and Thies, W. (2020). Exploring collection of sign language datasets: Privacy, participation, and model performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.

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

Gibet, S. (2018). Building french sign language motion capture corpora for signing avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Lee, S., Glasser, A., Dingman, B., Xia, Z., Metaxas, D., Neidle, C., and Huenerfauth, M. (2021). American sign language video anonymization to support online participation of deaf and hard of hearing users. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–13.

Loula, F., Prasad, S., Harber, K., and Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):210.

Lu, P. and Huenerfauth, M. (2010). Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 89–97.

McDermott, J. H. and Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940.

Portilla, J. and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70.

Tits, M. (2018). *Expert Gesture Analysis through Motion Capture using Statistical Modeling and Machine Learning*. Ph.D. thesis, Ph. D. Dissertation.

Troje, N. F., Westhoff, C., and Lavrov, M. (2005). Person identification from biological motion: Effects of structural and kinematic cues. *Perception & Psychophysics*, 67(4):667–675.

Zhang, Z. and Troje, N. F. (2005). View-independent person identification from human gait. *Neurocomputing*, 69(1-3):250–256.

8. Language Resource References

Benchiheub et al. (2020). *MOCAP1 corpus*. distributed via ORTOLANG: `https://hdl.handle.net/11403/mocap1/v1`, v1.