A new tool to facilitate prosodic analysis of motion capture data and a data-driven technique for the improvement of avatar motion

John McDonald[1], Rosalee Wolfe[1], Ronnie B. Wilbur[2], Robyn Moncrief[1], Evie Malaia[3], Sayuri Fujimoto[1], Souad Baowidan[1], Jessika Stec[1]

[1]DePaul University, Chicago, USA

[2]Purdue University, West Lafayette, USA

[3]Netherlands Institute for Advanced Study, the Netherlands

E-mail: jmcdonald@cs.depaul.edu, wolfe@cs.depaul.edu, wilbur@purdue.edu, rkelley5@mail.depaul.edu, evie.malaia@nias.knaw.nl, sfujimoto23@gmail.com, sbaowida@mail.depaul.edu, jessika.stec@gmail.com

Abstract

Researchers have been investigating the potential rewards of utilizing motion capture for linguistic analysis, but have encountered challenges when processing it. A significant problem is the nature of the data: along with the signal produced by the signer, it also contains noise. The first part of this paper is an exposition on the origins of noise and its relationship to motion capture data of signed utterances. The second part presents a tool, based on established mathematical principles, for removing or isolating noise to facilitate prosodic analysis. This tool yields surprising insights into a data-driven strategy for a parsimonious model of life-like appearance in a sparse key-frame avatar.

Keywords: motion capture analysis, sign language synthesis, avatar technology, noise

1. A simple case study as motivation

Noise is an unwanted modification to motion capture data that occurs during recording. The following example illustrates how noise poses barriers to the analysis of prosodic structure. Figure 1 is a time graph taken from a motion capture session [1]. It displays the y-coordinate (height) of the right wrist over a two-second period at the beginning of the sentence 'Newspaper said there was an awful storm in Florida where homes, cars, and trees were destroyed.' The first two seconds contain the signs 'NEWSPAPER READ'.

Although the signal looks smooth to the casual observer, problems arise when using the data to compute changes in speed as a precursor to examining the prosody of an utterance.

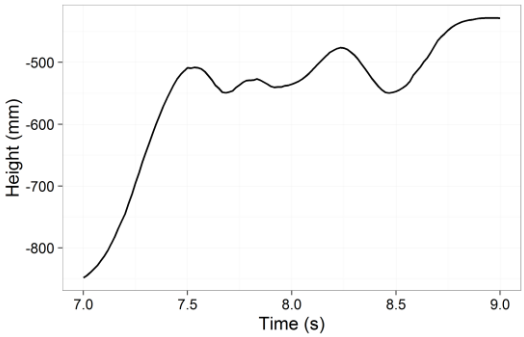

Figure 1: Height information for a right wrist marker.

Determining changes in speed is a two-step process. The first computes the speed from the marker's position data using a central difference approximation for the derivative:

$$s = \frac{dp}{dt} = \left| \frac{p_{i+1} - p_{i-1}}{2\Delta t} \right|$$

Figure 2 is a graph of the wrist marker's speed. The curve contains many small spikes which are due to the noise contained in the original position data.

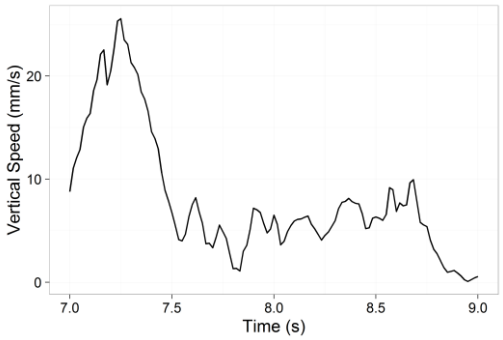

Figure 2: Speed of right wrist.

The second step computes the change in speed, which is essential for studying prosody:

$$\frac{ds}{dt} = \frac{s_{i+1} - s_{i-1}}{2\Delta t}$$

Figure 3 is a graph of the result. The spikes are even larger and dominate the curve. This jagged curve gives the impression of jerky motion, but the original position graph in Figure 1 reflects the smoothly flowing discourse of a fluent signer as confirmed in the original video.

The noise that was barely perceptible in Figure 1 has been magnified to the point where it is difficult to use visual inspection to identify any aspect in the prosodic structure of the utterance. From this example, it is clear that the motion capture data contains noise, but the question

remains as to its origins and severity. Effective analysis requires its isolation and/or removal.

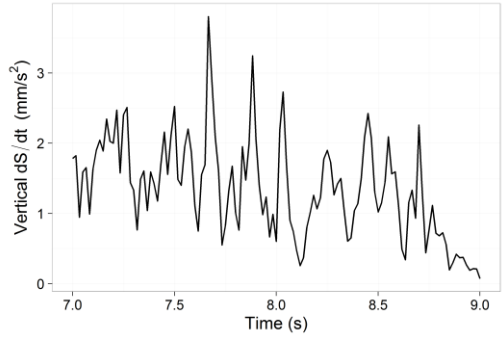

Figure 3: Change in right wrist speed

2. Fundamentals and terminology of signal processing

This section takes the form of a brief tutorial, outlining the principles of signal processing used to clean a motion capture data stream. These principles are applicable to the analysis of any time series data, including motion capture. The interested reader can find a more in-depth treatment in [2].

Several important concepts of signal processing can be analyzed from an idealized production of the word BICYCLE in American Sign Language (ASL). In this sign, the height of the right hand oscillates vertically in a regular manner similar to the idealized graph shown in Figure 4. Since the horizontal axis of this graph is time, this plot is said to be in the *time domain*.

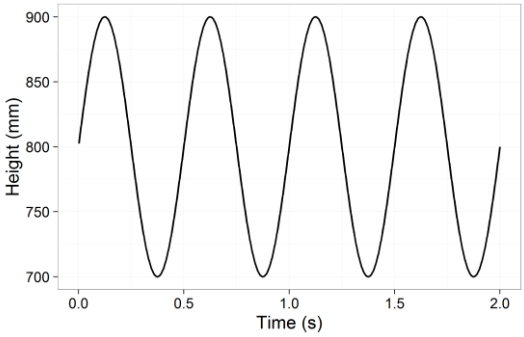

Figure 4: Height data of a wrist from an idealized production of BICYCLE.

The size of the oscillation is called the *amplitude* of the signal, whereas the speed at which the hand moves through the oscillation is its *frequency*. Amplitude is measured in units such as millimeters (mm), and frequency is measured in *cycles per second* also known as Hertz (Hz).

Unfortunately, the signal is rarely as simple as in Figure 4. Returning to Figure 1, the oscillations in the graph show variation in both their length and size. Thus, these oscillations change in both amplitude and frequency over the course of the phrase. To analyze more complicated signals, we need the Fourier transform [3], which decomposes a signal into a collection of contributing pure oscillations. Figure 5 shows a density plot, analogous to a histogram, of all the oscillations present in the signal from Figure 1. This plot is called the signal's *spectrum* in the *frequency domain*, since it displays the strengths of the signal's oscillations at various frequencies, which are shown on the horizontal axis.

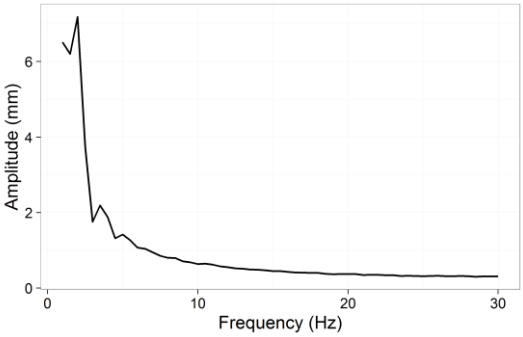

Figure 5: Frequency spectrum of the wrist height during the first 2 seconds of the phrase

This spectrum was constructed with a Fourier transform on the original time-domain signal, and yields a list of amplitudes in the frequency domain, which we can then analyze and edit. As an example, we return to the question of noise. Figure 6 contains a plot of the signer's right wrist height while standing still with arms raised in a calibration posture.

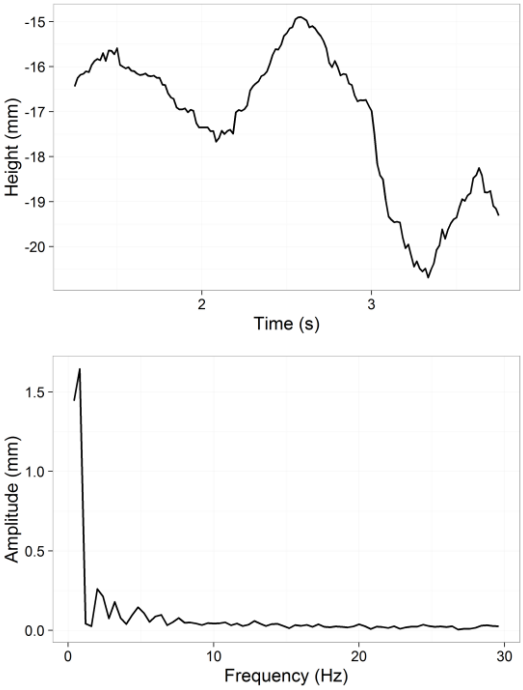

Figure 6: Time and frequency domain plots of signer with arms up

An analysis of the right wrist height and its resulting spectrum yields one main low-frequency oscillation with a spread of smaller amplitudes at higher frequencies. These are very fast, but tiny oscillations around a slow variation of the wrist height that occurs as the signer attempts to hold still.

Returning to the motion in Figure 5, we see a more complicated profile with a main high amplitude signal at

low frequencies and then a smooth falloff in amplitude at higher frequencies. Their small amplitudes indicate that these fast oscillations contribute little to the signal. It is this *noise* that software needs to remove before meaningful analysis can be performed.

For our purposes, removing unwanted high frequencies will not alter the main signing signal. We do this by means of a *low-pass filter*, which sets all the frequency amplitudes above a certain threshold to zero. After the suppression of these amplitudes, we can recover the cleaned signal by inverting the Fourier transform, yielding a smoother trajectory for the wrist. The cleaned signal will rarely deviate from the original by more than a fraction of a millimeter. In our study, over 99% of the samples deviated by less than a millimeter.

3. Analyzing Noise in Sign Language Motion Corpora

This section discusses practical considerations for determining which frequencies are relevant to linguistic research and which can be safely considered as noise. Figure 7 contains a conceptual diagram of a spectrum for a coordinate value of a position marker in the frequency domain. The vertical axis is amplitude and the horizontal axis is frequency.

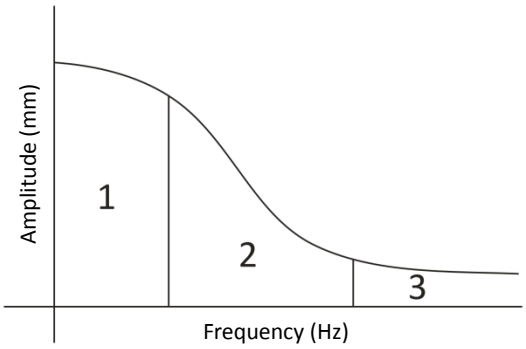

Figure 7: Conceptual regions of positional data graphed in the frequency domain.

The frequency spectrum in this diagram is divided into three sections which have different impacts on sign analysis. We begin with the region marked "3", representing frequencies above 12 Hz. According to [4], the muscles in the human body cannot create oscillations faster than 10-12 Hz, and so the frequencies in this region can thus be seen as noise attributable to fluctuations in the recording technology. These frequencies can safely be eliminated before performing further analysis of the signal.

Frequencies slower than 10-12Hz, in regions 1 and 2, may be produced by human motion. However, not all such frequencies of motion have linguistic meaning for sign language. This can be clearly seen by looking at the types of motion that the human body produces in sign discourse and the oscillations of parts of a signer's body involved in such motion. On the slower end of the scale, oscillations on hip markers correspond to such linguistic processes as *role shift*. Due to the sheer mass involved in moving the human torso, these motions will have lower frequencies of no more than 0.5 Hz. In contrast, fingers being of much lower mass and smaller movements, are capable of higher frequencies, such as the motion displayed in fingerspelling or in internal movement such as trilling (WAIT, FINGERSPELL), but even here the cutoff is no more than 4 Hz as can be seen in analyses of finger spelling rates[5].

Thus, the region in the diagram marked "1" contains the main low frequency movements generated by sign language production. The cutoff for this region will depend on a marker's placement, with lower frequencies for markers on the trunk of the body and higher frequencies at more distal markers. Table 1 gives a set of empirically-determined frequency cutoffs for intermediate markers. These limits are deliberately conservative to assure that no aspect of a human linguistic utterance is being compromised.

Joint	Frequency (Hz)
Hips	0.25
Waist	0.5
Upper spine	0.5
Neck	1.0
Shoulders	1.0
Elbow	2.0
Wrist	2.0

Table 1: Frequency cut offs for selected markers.

For linguistic analysis, we can clean the position data by converting it to the frequency domain, setting the amplitude of the frequencies in regions 2 and 3 to zero, and using the modified spectrum to reconstitute the marker's position in the time domain via an inverse Fourier transform. From the cleaned data, we proceed with the calculations for speed and speed change. The resulting graphs shown in Figure 8 do not exhibit the spikes seen in Figures 2 and 3.

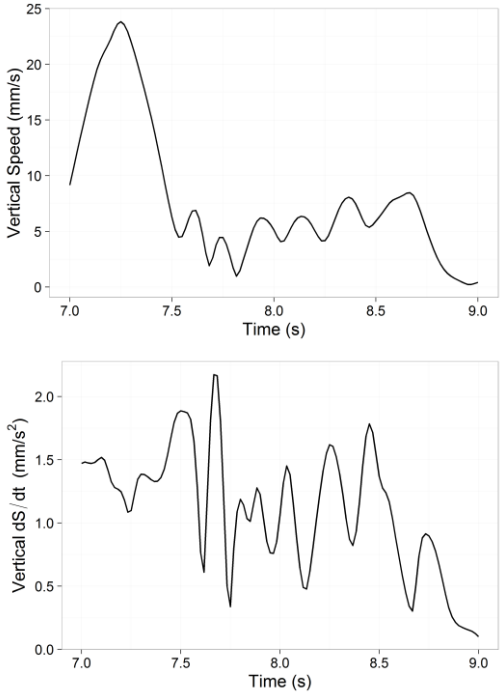

Figure 8: Speed and change of speed computed with cleaned position data

4. A new tool

To aid in isolating or removing noise from motion capture data, we created a software suite called SignCleaner to aid in the signal processing of motion capture data of signed utterances. The system accepts HTR, a common format of motion capture data [6] and can accommodate any number of markers. The suite is available for download at http://tinyurl.com/jfysn2t and consists of two parts. The first part is a C# application that translates HTR data into a comma delimited (.csv) file compatible with the R statistical computation environment [7]. The second part is a collection of R scripts that perform the following:

- Removing noise (cleaning) marker data using a Butterworth filter, which is based on a Fourier transform [8]. A Butterworth filter tapers the attenuation of the frequencies being removed for a highly smooth result. Researchers can adjust the frequency cutoffs to best accommodate their analyses.
- Computing speed and change of speed for each marker. Since these are scalar metrics, they lend themselves to easy visualization in the time domain.
- Visualizing the data to facilitate inspection for patterns or trends.
- Exporting the position, speed and speed change of markers as a CSV file, suitable for use in ELAN [9].

The tool has been validated on a subset of the Wilbur corpus [10], consisting of 58 markers with 9400 data points per marker. Figure 9 shows a screen shot of an ELAN session, showing a segment of the speed and change of speed of the right wrist sensor. For comparison, both measures are computed with the cleaned position data and the original, uncleaned data. The lighter curves in each track show the results from the original noisy position, whereas the darker curves are computed from the cleaned data.

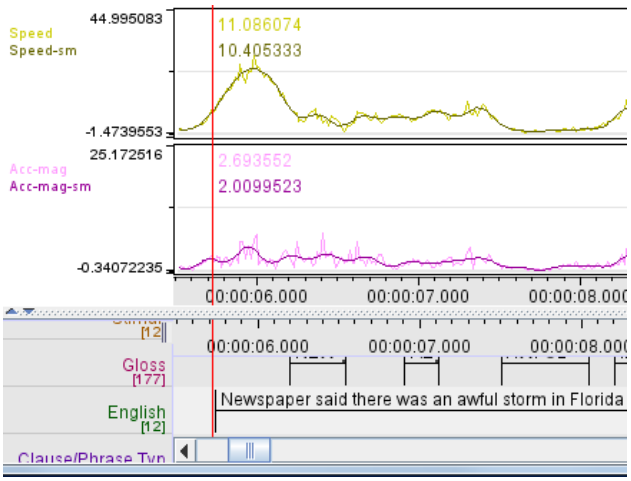

Figure 9: Elan interface for motion plot analysis

5. A novel finding and its application to avatar technology

Our previous discussion of Figure 7 did not consider the entire spectrum, so we return to it now. From the diagram, we know that we want to eliminate the frequencies in region 3 as they are noise introduced by the recording technology. Further we want to retain the frequencies in region 1 for linguistic analysis. This leaves region 2, which contains frequencies that are not of linguistic significance, but are none the less created by a human while producing signed utterances. From the perspective of linguistic analysis, this is noise, but from the perspective of avatar technology, this is valuable information for enlivening an avatar.

In order to create the illusion of life, avatars must continue to move, even when a signed discourse has concluded. A living human body is never completely still, even when at rest, and the human mind and visual sense are highly attuned to expect this dynamic. An avatar at rest needs to continually display subtle movements to avoid being perceived as a static image. This is a particular challenge for sparse-key animation systems [11]

In entertainment technology, two common techniques used to maintain the dynamics of an avatar are

- the manual adjustment of motion curves by an animation artist [12], and
- the introduction of Perlin noise.

Since hand animation is time-consuming and expensive, Perlin noise is preferred because it can be automated [13]. Perlin noise can be tuned to a specific set of frequencies [14] and is therefore ideal for this situation. We can tune this type of noise so that it primarily contains frequencies in region 2, the enlivening frequencies, and these will be perceptible in the finished animation. Figure 10 shows the frequency spectrum for a version of Perlin noise tuned to roughly match the three regions of Figure 7.

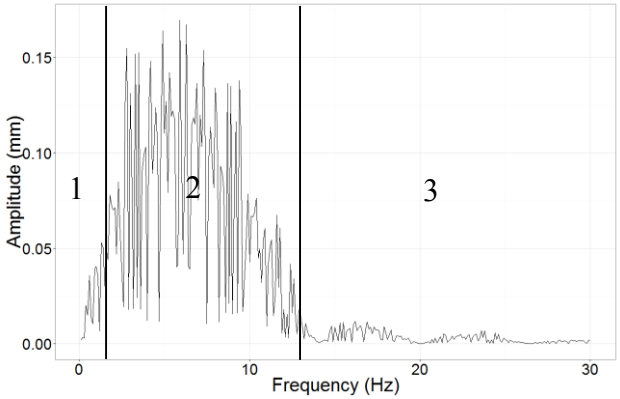

Figure 10: Spectrum of Perlin noise

The frequencies in this plot are essentially bounded on the right, and so there are very few high frequencies corresponding to region 3. In addition, the amplitudes of its low frequencies in region 1 are small enough so that the addition of this noise will not interfere with any intended animations such as a signed utterance. Since the range of frequencies is bounded on both the lower and upper ends, it corresponds nicely with enlivening region 2 of Figure 7.

Traditionally, Perlin noise is only applied in situations

where the avatar has otherwise stopped moving, however an abrupt transition to Perlin noise is incompatible with the high fidelity motion required to make avatar signing easily legible. Attempts to gradually introduce Perlin noise do not improve the problem, and can introduce jarring discontinuities in the motion.

6. An insight from motion capture data

A *heatmap* facilitates further exploration the presence of noise in the motion capture data by visualizing the relationship of frequency and amplitude with time in the signal. Figure 11 displays a heatmap of the amplitude/frequency profile over an entire recording session computed using a sliding discrete Fourier transform [15]. In this visualization, the *x*-axis displays the frequency, the *y*-axis displays time, and the amplitude is displayed as a grayscale intensity with darker intensities representing higher amplitudes. The regions labeled in this figure correspond to regions in Figure 7. For frequencies in region 3 that are greater than 12Hz, the noise is nearly constant over the entire time range. This is to be expected since this noise does not come from human movement, but rather from the recording equipment itself.

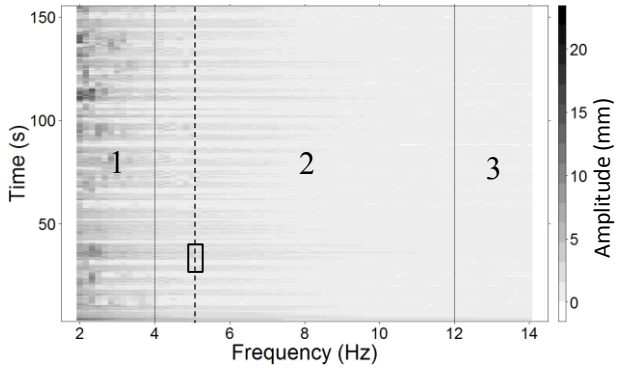

Figure 11: Heatmap of amplitude vs frequency and time for right wrist height

The frequencies in region 2 are too high to warrant linguistic analysis, but are still produced by a human signer. The heatmap demonstrates that these frequencies are present throughout the entire discourse, whether the signer is producing utterances or is at rest. From a linguist's perspective, this is noise and can safely be ignored, but from an animator's perspective, region 2 frequencies are actually invaluable, as they can be used to enliven the avatar. These data inform us that these frequencies must be present whenever an avatar is signing or is at rest. Observers do not perceive these frequencies as noise during signing, since the frequencies of the signed utterances have comparatively higher amplitude. These high-amplitude motions produced by signing overwhelm the subtle changes created by the lower amplitude frequencies from region 2.

To further investigate the relationship between noise and signing, we examine a representative clip of the height of the right wrist marker during two sentences which begin and end with the signer at rest. We will focus on a frequency of 5Hz which lies in the enlivening region of the heat map.

A vertical slice of the heatmap at 5Hz, corresponding to the dotted line in Figure 11, can be plotted with time on the *x*-axis and the amplitude at 5Hz on the *y*-axis. Figure 12 shows the graph of the portion of this signal corresponding to the small rectangle in Figure 11. Active signing in this segment occurs between times 24 and 30 seconds. The signer is at rest at the onset and conclusion of the segment.

The conventional expectation would be that the amplitudes for this particular frequency should be lower while the person is signing. Yet in this example we find exactly the opposite. Counterintuitive as it is, the enlivening frequencies are not just present, but actually *increase* in amplitude in the center of this graph, during which the signer is actively producing utterances. So, when adding noise to enliven an avatar, we should not suppress or turn off that noise when the avatar is signing. Figures 11 and 12 thus provide additional evidence that we should apply these enlivening frequencies throughout an avatar's signing.

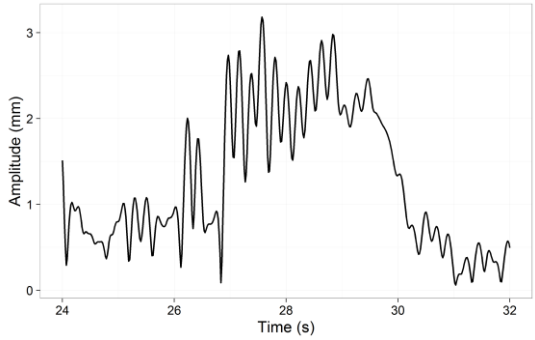

Figure 12: Amplitude of wrist height at 5Hz for two sentences

7. Implementation

To add enlivening frequencies to the avatar, we apply Perlin noise generators to each joint using the frequency ranges dictated by region 2. The generators run continually, and independently, of any utterances produced by the avatar. The exception to this is the blinking action of the avatar's eyelids. Blinking is a discrete movement that needs to be controlled with a separate mechanism which is outside the scope of this paper [15] [16].

There is one additional consideration required when setting up the Perlin noise generators, as they also require knowledge of amplitude. This information is easily obtainable from the spectrum of each marker and is summarized in Table 2. Because our avatar requires angle data for its joint rotations, we capitalize on the fact that $sin(\theta) = \theta$ for small θ, thus easily converting the positional data into rotational data.

The Perlin noise generators add a modest computational cost, but if the avatar is being used in an environment where computing resources are limited, then implementing a single generator on the hips is an effective choice as the hips will transmit subtle motion, albeit coordinated, to the rest of the avatar's skeleton, even in the absence of noise on the other joints. [17].

Joint	Amplitude (degrees)
Hips	6.37×10^{-3}
Waist	4.78×10^{-3}
Upper spine	4.78×10^{-3}
Neck	2.39×10^{-3}
Shoulders	2.39×10^{-3}
Elbow	2.39×10^{-3}
Wrist	2.39×10^{-3}

Table 2: Amplitudes (noise strengths) for Perlin noise generators.

8. Results

To test this approach, we applied Perlin noise generators to all the joints in the avatar's spinal column (hips, waist, upper spine, and neck) and arms (shoulder, elbow, and wrist). More distal joints were given noise with lower amplitudes and higher frequencies as indicated in Tables 1 and 2. The generators are active throughout the entire animation, regardless of whether the avatar is signing or not.

The reference http://tinyurl.com/zzl8btc is a link to a video demonstrating the effect. The video contains a side-by-side comparison of animations with and without Perlin noise generators. The animation on the left has no noise, while the one on the right has noise applied to all joints previously mentioned. When at rest, the figure on the left has the appearance of a static photograph, whereas the figure on the right continues moving subtly. The noise does not interfere with the portrayal of the signed utterances.

This approach is well accepted by test participants who view and rate our avatar's utterances for clarity and naturalness. In a developing a mathematical model for role shift as reported in [18], Deaf participants fluent in ASL viewed and rated animations that incorporated this livening method. A majority of the participants rated clarity as either "clear" or "very clear" on a 5-point Likert scale. A follow-up study [19] yielded similar results. Clarity was a particularly important measure here, because it tested whether noise was interfering with the avatar's signing. The results indicate that applying noise to an avatar's joints, with frequencies and amplitudes appropriately tuned according to the results of the study of motion capture data, are effective in enlivening an avatar without impeding the avatar's ability to communicate.

9. Future work

We look forward to testing the scalability of SignCleaner by applying it on larger corpora. We also plan to use it for its original intended purpose of prosodic analysis. In addition, we will add the ability to import other motion capture formats.

Acknowledgment

We extend our gratitude to Eleni Efthimiou for collegial discussions on the challenges and nature of the application of noise to an avatar system.

10. References

1 Wilbur, Ronnie B. and Malaia, Evgenia. A new technique for analyzing narrative prosodic effects in sign languages using motion capture technology. In (eds.), A. Huebel & M. Steinbach, ed., *Linguistic foundations of narration in spoken and sign languages*. John Benjamins, Amsterdam, In Press.

2 Smith, Steven. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, Santa Clara, CA, 2011.

3 Duhamel, Pierre and Vetterli, Martin. Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing* (1990), 259-299.

4 Marshall, John and Walsh, E. Geoffrey. Physiological Tremor. *Journal of Neurology, Neurosurgery, and Psychiatry* (1956), 260-267.

5 Quinto-Pozos, David. Rates of fingerspelling in American Sign Language. In *Theoretical Issues in Sign Language Research* (West Lafayette, IN 2010).

6 Parent, Rick, Ebert, D. S., Gould, D. et al. *Computer animation complete: all-in-one: learn motion capture, characteristic, point-based, and Maya winning techniques*. Morgan Kaufman, Burlington, VT, 2009.

7 R CORE TEAM. *R Language Definition*. 2000. ftp://155.232.191.133/cran/doc/manuals/r-devel/R-lang.pdf. Accessed 2016-1-18.

8 Hong, Jia-Shen G. and Lancaster, Michael J. *Microstrip filters for RF/microwave applications*. John Wiley & Sons, Hoboken, NJ, 2004.

9 Crasborn, Onno, Sloetjes, Han, Auer, Eric, and Wittenburg, Peter. Combining video and numeric data in the analysis of sign languages within the ELAN annotation software. In *Proc. LREC 2006 workshop on representation & processing of sign language* (Paris 2006), ELRA, 82-87.

10 Wilbur, Ronnie, Gokgoz, Kadir, Shay, Robin, and Martínez, Aleix. Mechanics and Issues in Making ASL Databases Available. In *Building sign language corpora in North America* (Washington, DC 2011).

11 Perlin, Ken. A system for scripting interactive actors in virtual worlds. In *Proceedings of ACM SIGGRAPH 96* (New Orleans 1996), Association for Computing Machinery, 205-216.

12 Gleicher, Michael. Retargetting motion to new characters. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), 33-42.

13 Perlin, Ken. Real time responsive animation with personality. *IEEE Transactions on Visualization and*

Computer Graphics (1995), 5-15.

14 Lagae, Ares, Lefebvre, Sylvain, Cook, Rob et al. A survey of procedural noise functions. *Computer Graphics Forum* (2010), 2579-2600.

15 Jacobsen, Eric and Lyons, Richard. The Sliding DFT. *Signal Processing Magazine, IEEE*, 20, 2 (March 2003), 74-80.

16 Baker, Charlotte and Padden, Carol. Focusing on the Nonmanual Components of American Sign Language. In Siple, Patricia, ed., *Understanding Language through Sign Language Research*. Academic Press, New York, 1978.

17 Wilbur, Ronnie B. Eyeblinks and ASL phrase structure. *Sign Language Studies* (1994), 221-240.

18 McDonald, John, Wolfe, Rosalee, Schnepp, Jerry et al. An Automated Technique for Real-Time Production of Lifelike Animations of American Sign Language. *Journal of Universal Access in the Information Society* (2015), 1-16.

19 Schnepp, Jerry, Wolfe, Rosalee, McDonald, John, and Toro, Jorge. Generating Co-occurring Facial Nonmanual Signals in Synthesized American Sign Language. In *Eight International Conference on Computer Graphics Theory and Applications* (Barcelona 2013), 407-416.

20 McDonald, John, Wolfe, Rosalee, Moncrief, Robyn, and Baowidan, Souad. A computational model of role shift to support the synthesis of signed language. In *Theoretical Issues of Sign Language Research (TISLR)* (Melbourne, VIC 2016).