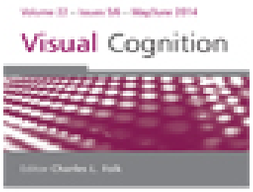

Visual Cognition

ISSN: 1350-6285 (Print) 1464-0716 (Online) Journal homepage: http://www.tandfonline.com/loi/pvis20

Assessment of information content in visual signal: analysis of optical flow fractal complexity

Evie Malaia, Joshua D. Borneman & Ronnie B. Wilbur

To cite this article: Evie Malaia, Joshua D. Borneman & Ronnie B. Wilbur (2016): Assessment of information content in visual signal: analysis of optical flow fractal complexity, Visual Cognition, DOI: 10.1080/13506285.2016.1225142

To link to this article: http://dx.doi.org/10.1080/13506285.2016.1225142

VISUAL COGNITION, 2016
http://dx.doi.org/10.1080/13506285.2016.1225142

Routledge
Taylor & Francis Group

Assessment of information content in visual signal: analysis of optical flow fractal complexity

Evie Malaia[a], Joshua D. Borneman[b] and Ronnie B. Wilbur[b]

[a]Netherlands Institute for Advanced Study, Meijboomlaan 1, 2242 PR Wassenaar, The Netherlands; [b]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN 47907, USA

ABSTRACT

We make a first attempt at distinguishing an information-carrying visual signal by comparing visual characteristics of American Sign Language to everyday human motion, to identify what clues might be available in one but not in the other. The comparison indicated significantly higher fractal complexity in sign language across tested frequency bands (0.01–15 Hz), as compared to everyday motion. A comparison of our results with other work showing high fractal complexity in the speech signal allows us to suggest the underlying properties of linguistic signals which allow babies to "tune into" a specific channel, or modality, during language acquisition.

ARTICLE HISTORY
Received 17 May 2016
Accepted 31 July 2016

KEYWORDS
ASL; sign language; fractal complexity; motion; optical flow

Approaches to the puzzle of acquisition of language have tended to focus on issues related to the segmentation of the auditory stream using statistical, prosodic, and social cues (Johnson, Seidl, & Tyler, 2014; Seidl, Tincoff, Baker, & Cristia, 2015). However, experimental evidence shows that babies can identify the information-carrying channel during the language acquisition period even if it is visual: for example, hearing babies of deaf parents try to "babble" using their hands (Petitto, Holowka, & Sergio, 2001).[1] To our knowledge, the question of how the brain recognizes an information-rich linguistic signal regardless of its physical domain has never been addressed. How does the language-ready brain of a deaf baby with no prior auditory exposure recognize a linguistic component in the visual input? We suggest that the signal must stand out from the surrounding background in a way that is identifiable by the human neural system.

We take the first step toward quantifying possible universal properties of the linguistic signal by approaching it from the perspective of an essential function of communication: information transfer. The standard quantifiable measure of information is entropy: the uncertainty involved in predicting the next data point in a time series (Shannon, 1948). A signal's spectral fractal complexity is an abstract quantitative measure based on Shannon entropy, which captures potential information-carrying capacity of a channel. In the auditory domain, where the linguistic signal is easily described as a series of sounds with specific characteristics, languages of the world are described as having a modulation spectra of moderate fractal complexity (Singh & Theunissen, 2003). However, the underlying properties of the visual linguistic signal allowing babies to "tune into" a specific channel/modality have not been described and, with respect to sign languages, the issue is made even more interesting given the added problem of non-linearity (multi-channel representation) in sign languages.

Recently documented phonotactic and grammatical roles of motion in the linguistic systems of unrelated sign languages (Malaia, Ranaweera, Wilbur, & Talavage, 2012; Malaia & Wilbur, 2012; Malaia, Wilbur, & Milković, 2013) suggest that humans who are exposed to linguistic visual signals develop an ability to produce and analyse complex signals in the visual domain. In this work, we tested the hypothesis that the visual linguistic signal (American Sign Language) has higher information-carrying capacity than everyday motion across the spectrum of visible frequencies, using optical flow analysis.

Materials and methods

Our comparison is based on a visual analysis of existing videos of two types of motion. The first type is

CONTACT Evie Malaia evie.malaia@nias.knaw.nl Netherlands Institute for Advanced Study, Meijboomlaan 1, 2242 PR Wassenaar, The Netherlands

commonly referred to as everyday motion or gesture, and derives from movements made by humans while conducting their normal routines. These videos were stimuli produced for motion-event boundary experiments conducted by Zacks, Kumar, Abrams, and Mehta (2009) and were kindly provided to us for this analysis. Included in these videos were everyday human activities, such as laundry folding, video game assembly, and Lego construction (from Zacks et al., 2009). The idea underlying these activities is that they consist of sub-events, such as folding a shirt sleeve, that together constitute a larger (or macro) event, such as a completely folded shirt or a stack of folded laundry. The motions produced by a single hearing actor are purposeful, sequential, and varied. Zacks et al. studied the correlations between actor movements and viewers' perceptions of where events ended as a test of their Event Segmentation Theory, which "proposes that everyday activity includes substantial sequential dependency" (p. 202). In addition, viewers may also use cues from the actor's facial expressions or eye-gaze, or from interaction with objects in the environment.

The second type of video contains narratives in American Sign Language that were created for prior studies of our own (Malaia et al., 2012; Malaia, Borneman, & Wilbur, 2008; Malaia & Wilbur, 2012). The signed motions produced by a single deaf signer are meaningful sentences consisting of sequences of various signs, and are thus in many ways comparable to the everyday motion videos. However, sign languages make deliberate and meaningful use of the space in front of the signer, as well as various articulators of the face (eyes, brows, mouth) and head and body positions (Wilbur, 2000). These additional meaning-carrying units are co-articulated with the signs themselves and increase the number of cues transmitted simultaneously. For example, questions requiring a "yes" or "no" response are made with raised eyebrows, whereas those requiring a contentful answer (to "who/what/when") used lowered brows. Spatial layout can indicate which person being talked about is the subject of the sentence, even if the person being talked about is not present in the conversation. It could be argued that these additional cues are much like those accompanying everyday activities, that is, facial expression, eye-gaze, or interacting with objects. Thus our comparison is aimed at identifying whether the signing signal is visually distinctive from the everyday motion in a quantitative way.[2]

Empirically, information transfer is implicit in sign language data: the videos used for analysis were both informative and comprehensible to signers. As to biological motion, it is also informative to an extent that humans are capable of segmenting it both behaviourally and perceptually (neutrally), with high cross-correlation among the viewers (Malaia, 2014; Noble et al., 2014; Zacks et al., 2009). Humans, regardless of prior exposure to sign language, are also capable of segmenting sign language narratives and identifying event boundaries in signed discourse[3] (Fenlon, Denmark, Campbell, & Woll, 2008; Strickland et al., 2015; Wilbur & Malaia, 2008).

The signing and non-signing videos (20 of signing and 40 of everyday motion) contained 1350 frames (30 fps × 45 sec) and had been recorded at 768 × 512 pixels. We converted them to greyscale colour. Given their origins as stimuli for other experiments, each video contained a participant in front of a static, uniform background. Optical flow of a video frame is the distribution of apparent velocities of objects in an image; that is, a velocity vector (in pixels/frame) is found for each pixel, based on how fast and in which direction, the feature shown in that pixel has moved from the frame before.

To reduce potential variations in motion velocity magnitude (ignoring direction) across the data set due to differences in original camera field-of-view and distance (resulting in differences in the relative size of the person), the videos were scaled to achieve uniform participant size. This was done by measuring a common reference on all videos (the participant's upper arm length), after which the videos were resized so that this participant reference was the same length (in pixels) across the entire data set. The upper arm was selected because it was most often perpendicular to the camera axis. Potential references such as hands and forearms were considered, but proved problematic due to rotation directly towards or away from the camera, resulting in an artificial reduction in size on the video frame, which would affect proper scaling. Although comparisons were done with unscaled videos, which determined that the relative changes in velocity, or the frequency components of the optical flow, remained unchanged, scaling the videos allows us to obtain similar values for optical flow velocity

magnitude (it does not affect direction) across all videos for each step of the analysis. Scaling for consistent participant size results in videos of magnified or reduced resolutions compared to each original video, therefore after scaling the videos were padded and/or cropped so that all videos were the same resolution; this creates a consistent spatial frequency sampling across the entire data set. The maximum extent of participant motion from the centre of the video frame was identified for each video, and cropping was done selectively so that no motion information was lost. Where padding was needed in order to match the dimensions of the entire data set, a single colour border (grey value equal to the average of the entire video frame) was added to the video frame. The resulting final data set of 60 videos were all 500×301, greyscale, 30 fps, 45 second duration, with a consistent participant size. An example frame is shown in Figure 1.

Optical flow for each video was determined using the Mathwork's MATLAB vision toolbox optical flow function. This function was utilized to compare each video frame with the prior frame and, using a Horn-Schunck method (Horn & Schunck, 1981), an output matrix of size equal to the input video frame was calculated. Each element of the matrix identifies the magnitude of optical flow velocity (pixels per frame) between the two frames for each corresponding pixel in the video. This approach collects the total motion between each video frame regardless of the object; having a static background is important since all motion is relevant and measured.

The resulting optical flow matrix for each frame was then reshaped to a single 150,500 (500×301) element vector and then the pixel velocities were binned into 200 bins from 0 to 0.4 pixels/frame, which represents the minimum and maximum velocities across the entire video set. This results in a matrix containing the magnitude of optical flow (velocity) versus time where optical flow ranges from 0.0 to 0.4 pixels/frame (in 200 bins) and time ranges from 0 to 45 seconds (for 1350 frames).

Next, for each optical flow velocity, we look at the changes to that velocity over time, that is, at the frequency modulation of the optical flow signal. This is done by taking a one dimensional fast-Fourier transform on the optical flow vs. time vector. This gives a vector for the spectral density of the motion (optical flow) versus frequency component. Frequencies are defined from 0.01 to 15 Hz based on the video duration (45 seconds) and frame rates (30 Hz)

The optical flow spectral density was then analysed according to its fractal complexity. The function given in Equation (1) was fit, using an iterative nonlinear least squares method, to each frequency profile, where M is the magnitude of optical flow (non-directional), f is the frequency, α is a fitting parameter for the spectral density amplitude, and β is a fitting parameter for fractal complexity:

$$M(f) = \alpha / f^{\beta} \tag{1}$$

Thus, for each of the 200 optical flow velocity bins, an amplitude fitting variable and the fractal complexity

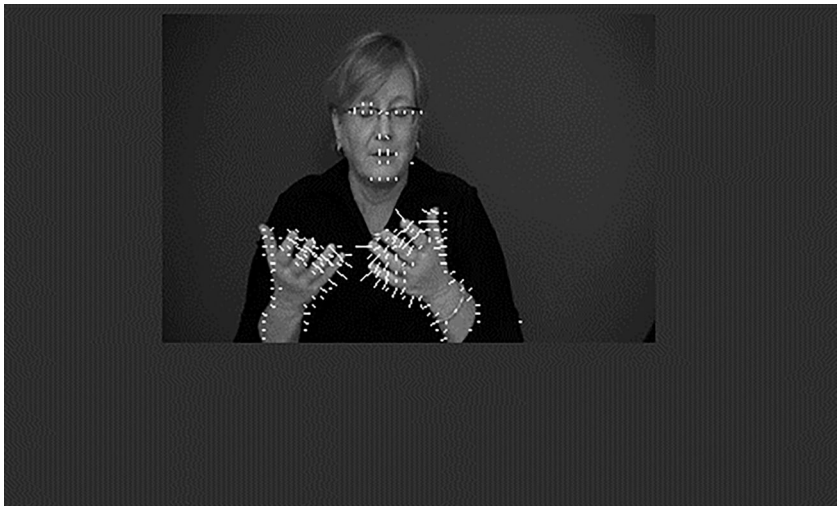

Figure 1. Example sign language video frame. Image has been scaled and cropped so that subject is uniformly located and sized across the entire video set. Arrows show motion vectors of optical flow.

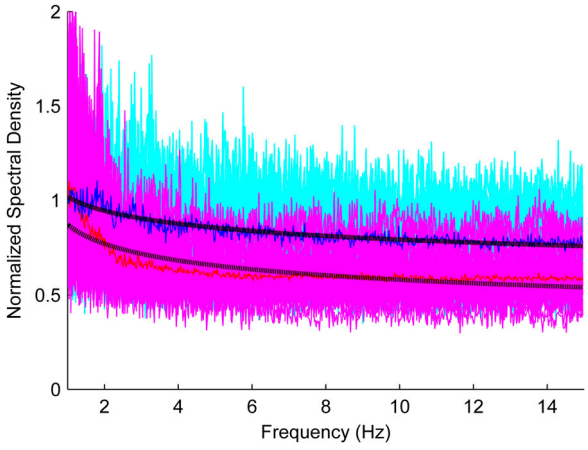

Figure 2. Example normalized spectral density for ASL (cyan/blue) and everyday motion (magenta/red). Cyan/magenta lines show raw data for Optical Flow between 0.20–0.25 px/sec, and the average for ASL (blue) and everyday motion (red) over all videos is also shown. Black lines show the respective fit according to Equation (1). Signing videos show greater fractal complexity.

parameter are determined to describe the dependence of optical flow spectral density on frequency. The spectral density of each video's optical flow is presented in Figure 2, showing spectral density versus frequency.

Results

We quantified the fractal complexity of motion by analysing the frequency profiles of optical flow for the two types of videos: (1) everyday activities, and (2) short ASL narratives. Spectral density amplitude and fractal complexity parameter were extracted with an average root mean square error (RMSE), across all fits, of 2%, with a global maximum RMSE of 3%. The extracted fractal complexity for a subset of optical flow values is presented in Figure 2, also showing normalized fractal complexity versus frequency.

Independent-samples *t*-test on the fractal complexity parameter (β) for each video indicated that on average, across the frequency range 0.01–15 Hz, the values of sign language videos were higher ($M = 0.271$, $SD = 0.143$) than those of everyday motion videos ($M = 0.364$, $SD = 0.124$), resulting in higher fractal complexity of optical flow across the tested frequency range in sign language. The difference was significant (t (57) for individual bins varied from −2 to −9, $p < .001$), and represented a medium-to-large size effect ($r = 0.25$ to 0.80).

Discussion

One of the fundamental goals of language research is the identification of signal properties distinguishing linguistic communication from other activities. Our results for ASL are comparable to that reported for the auditory domain, where world languages are described as complex systems following power law[4] spectral behaviour (Baken, 2010; Singh & Theunissen, 2003). Thus, as spoken language is distinct from surrounding environmental sounds, ASL is visually distinct from everyday activities despite their similarities in sequentiality. What we are suggesting, then, is that this distinctiveness, which we have characterized here in terms of fractal complexity/dimension, may be processed by the visual system and may serve as a marker to attract the attention of babies when signing is present in the baby's environment, because the information-transfer capacity of the signal fits with both perceptual processing and the neural computational system. It could be said that the brain is looking for, and receiving, a particular level of (linguistic) complexity in either the visual or the auditory domain. This "fit" of communicative signal properties to perceptual and neural systems underlies the subsequent production of language, first as infant babbling (vocal or manual) and later as recognizable language.

Our research contributes a piece to another puzzle as well. While speech has been shown to conform to a power law in its spectral complexity (Baken, 2010; Singh & Theunissen, 2003), it has remained unknown whether this is merely a parameter of verbal speech, given similar characteristics in zebra finch songs and higher frequency natural sounds (Singh & Theunissen, 2003), or whether it is a result of a fundamental underlying phenomenon. The finding that comparable complexity is found in both spoken and signed language allows us to suggest that power law spectral complexity is related to the communicative function, not merely the auditory domain.

Based on previous work identifying motion as a key component in syntax and semantics of sign languages (Brentari, 1998; Malaia et al., 2008), we characterized the information-carrying property of sign language in terms of fractal complexity of motion, based on mathematical analyses of information transfer between complex systems (West & Grigolini, 2010). The comparison indicated significantly higher fractal

complexity in sign language across tested frequency bands (0.01–15 Hz), as compared to everyday human motion. Interestingly, both everyday motion and sign language appeared to have a scale-free distribution of fractal complexity—a feature not unexpected in a biological system, but never previously documented for sign language. The current finding that it is visually distinct from everyday activity as determined from video is encouragement for continued pursuit of the motion kinematic analysis in sign languages.

Similarly, investigation of neuronal tuning shows that neurons in V1 area of the macaque brain are tuned to optimally respond to 1/f signal complexity in visual signals, as compared to $1/f^0$ or $1/f^2$ (Yu, Romero, & Lee, 2005), suggesting a fundamental biological basis for neural sensitivity to a specific range of fractal complexity in visual stimuli. General spatio-temporal structure of neural oscillations in the human brain also obey power-law scaling behaviour (Linkenkaer-Hansen, Nikouline, Palva, & Ilmoniemi, 2001). Interestingly, analyses of preferences for visual art has shown that in terms of aesthetics, humans also favour a specific complexity range (Taylor et al., 2005). The question of whether language as a communicative device overlaps with complexity ranges preferred for art in the respective domain (visual and auditory) will require further study. However, a quantitative approach to analysis of communicative signals can be a starting point for the development of more sophisticated methods of diagnostics for language acquisition and delay, both in the auditory and the visual modality. Further studies could match the stimuli across different dimensions of communicative/information transferring intent (using, for example, biological motion, co-speech gesture, pantomime, and a range of sign languages at various points of acquisition), as well as use visual signals of higher complexity ($\sim 1/f^2$) to investigate limitations and clinical usability of the proposed technique.

Notes

1. SL acquisition data showed a higher range of velocities, and thus potentially higher fractal complexity in hand motions of sign-language exposed infants of 6–12 months of age (Petitto et al., 2001).
2. Our current analysis seeks to capture information transfer in novel signed sentences much like those that would occur in a conversation, rather than the type of (rehearsed) 'sequence and repetition' information throughput characterized by Oulasvirta, Roos, Modig, and Leppänen (2013).
3. The ability to identify event boundaries in signed narratives does not render non-signers capable of understanding sign languages, or users of one sign language capable of understanding another sign language.
4. Power law is the relationship between two quantities, where one varies as a power of another. Power law relationships are characteristic of the mathematical class of complex systems, and have been shown to be relevant to, for example, frequency of words in a written text (Zipf, 1949).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by a EURIAS fellowship to EM.

ORCiD

Evie Malaia http://orcid.org/0000-0002-4700-0257
Ronnie B. Wilbur http://orcid.org/0000-0001-7081-9351

References

Baken, R. J. (2010). Irregularity of vocal period and amplitude: A first approach to the fractal analysis of voice. *Journal of Voice*, *4*(3), 185–197.

Brentari, D. (1998). *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Fenlon, J., Denmark, T., Campbell, R., & Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, *10*(2), 177–200.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*, 185–203.

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PloS One*, *9*(1), e83546. doi:10.1371/journal.pone.0083546

Linkenkaer-Hansen, K., Nikouline, V. V., Palva, J. M., & Ilmoniemi, R. J. (2001). Long-range temporal correlations and scaling behavior in human brain oscillations. *The Journal of Neuroscience*, *21*(4), 1370–1377.

Malaia, E. (2014). It still isn't over: Event boundaries in language and perception. *Language and Linguistics Compass*, *8*(3), 89–98.

Malaia, E., Borneman, J., & Wilbur, R. B. (2008). Analysis of ASL motion capture data towards identification of verb type. In *Proceedings of the 2008 conference on semantics in text processing* (pp. 155–164). Stroudsburg, PA: ACL.

Malaia, E., Ranaweera, R., Wilbur, R. B., & Talavage, T. M. (2012). Event segmentation in a visual language: Neural bases of

processing American Sign Language predicates. *Neuroimage*, *59*(4), 4094–4101.

Malaia, E., & Wilbur, R. B., (2012). Kinematic signatures of Telic and Atelic events in ASL predicates. *Language and Speech*, *55*(3), 407–421.

Malaia, E., Wilbur, R. B. & Milković, M. (2013). Kinematic parameters of signed verbs. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1677–1688.

Noble, K., Glowinski, D., Murphy, H., Jola, C., McAleer, P., Darshane, N., … Pollick, F. E. (2014). Event segmentation and biological motion perception in watching dance. *Art & Perception*, *2*(1–2), 59–74.

Oulasvirta, A., Roos, T., Modig, A., & Leppänen, L. (2013, April). *Information capacity of full-body movements*. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1289–1298). New York, NY: ACM.

Petitto, L. A., Holowka, S., & Sergio, L. E. (2001). Language rhythms in baby hand movements. *Nature*, *413*, 35–36.

Seidl, A., Tincoff, R., Baker, C., & Cristia, A. (2015). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental Science*, *18*(1), 155–164.

Shannon, C. E. (1948). Mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394–3411.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, *112*(19), 5968–5973.

Taylor, R. P., Spehar, B., Wise, J. A., Clifford, C. W., Newell, B. R., Hagerhall, C. M., … Martin, T. P. (2005). Perceptual and physiological responses to the visual complexity of fractal patterns. *Nonlinear Dynamics, Psychology, and Life Sciences*, *9*, 89–114.

West, B. J., & Grigolini, P. (2010). The living matter way to exchange information. *Medical Hypotheses*, *75*(6), 475–478.

Wilbur, R. B. (2000). Phonological and prosodic layering of nonmanuals in American Sign Language. In K. Emmorey & H. Lane (Eds.), *The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima* (pp. 213–241). Hillsdale, NJ: Lawrence Erlbaum.

Wilbur, R. B., & Malaia, E. (2008). Contributions of sign language research to gesture understanding: What can multimodal computational systems learn from sign language research. *International Journal of Semantic Computing*, *2*(1), 5–19.

Yu, Y., Romero, R., & Lee, T. S. (2005). Preference of sensory neural coding for 1/f signals. *Physical Review Letters*, *94*(10), 108103. doi:10.1103/PhysRevLett.94.108103

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort. An introduction to human ecology*. Cambridge, MA: Addison-Wesley.