OSCILLATORY GESTURES AND DISCOURSE

Francis Quek and Yingen Xiong

CSE, Wright State University
3640 Colonel Hwy, Dayton, OH 45435

ABSTRACT

Gesture and speech are part of a single human language system. They are co-expressive and complementary channels in the act of speaking. While speech carries the major load of symbolic presentation, gesture provides the imagistic content. Proceeding from the established cotemporality of gesture and speech, we discuss our work on oscillatory gestures and speech. We present our wavelet-based approach in gestural oscillation extraction as geodesic ridges in frequency-time space. We motivate the potential of such computational cross-modal language analysis by performing a micro analysis of a video dataset in which a subject describes her living space. We demonstrate the ability of our algorithm to extract gestural oscillations and show how oscillatory gestures reveal portions of the discourse structure.

1. INTRODUCTION

Gesture and speech are part of a singular language system, and as such, they are co-expressive and complementary [1]. Human multimodal language performance proceeds from the same 'idea units', move through the brain, and eventually result in the respective motor activities (speech is ultimately motor with synchronized lung, laryngeal, jaw, etc. activity). Hence, gesture and speech cohere temporally at the level of discourse 'idea pulses' and reveal the structure of discourse [2, 3, 4]. Speech necessarily encodes these units syntactically and symbolically, and gesture provides an imagistic, analog representation. While the formation of idea pulses is not observable, they are evidenced by the cohesive speech and gestural representations. We argue that the imagism in gestures are borne, not by 'whole gesture' performances, but by certain gestural features (e.g. the use of spatial loci of the hands to index conceptual space [5, 6]). This is analogous to series of peaks in a mountain range that inform us that they were formed by a common underlying (unobservable) process because they share some geological characteristic (even if there are peaks of heterogeneous origins that punctuate the range).

Obviously not all hand gesture features have equal potential for bearing the imagistic content. The abduction angle of the left little finger is probably of unimportant. The question, then, is what gestural features have greater power for discourse structure recovery [4]. In this paper, we investigate the oscillatory behavior of the speaker's hands. We observe that repetitive gestures are common, and that they seem to mark cohesive discourse pieces [1] (a simple mental experiment in which the reader attempts to break speech phrases out of synchrony with the beginning and termination of oscillatory gestures will be instructive). We present a wavelet-based approach for extracting oscillatory behavior, and its application in the microanalysis of the discourse structure of subject describing her living quarters to demonstrate the potential of oscillatory gestures as the 'material carrier' of the discourse content [7].

2. WAVELET-BASED OSCILLATION DETECTION

We apply a continuous wavelet transform (CWT) approach to detect the oscillatory motion of the subject's hands from their motion traces. The advantage of this approach over such frequency-domain approaches as the fast fourier transform (FFT) is that the CWT produces an explicit frequency-time space from which both the temporal extent and the oscillation frequency may be detected simultaneously [8]. The application of a windowed FFT addresses this problem somewhat, but it begs the question of the size of the window to be applied. At high frequencies, smaller time windows may be applied, but lower frequencies require larger windows with the attendant loss of temporal resolution. It also necessitates the application of multiple FFTs over the same signal. The CWT does not suffer from these shortcomings. We introduce two enhancements to CWT frequency-time space analysis. First, we apply a continuous varying frequency ridge detector at each time instant. Second, we demonstrate the ability to extract multiple simultaneous frequency traces. This is essential for gesture analysis since gesticular motion exhibits such complex multi-frequency behavior.

This research has been supported by the U.S. NSF STIMULATE program, Grant No. IRI-9618887, "Gesture, Speech, and Gaze in Discourse Segmentation", and the NSF KDI program, Grant No. BCS-9980054, "Cross-Modal Analysis of Signal Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research".

0-7803-7663-3/03/$17.00 ©2003 IEEE

A CWT is defined as

$$C_{a,b} = \int_R s(t)\frac{1}{\sqrt{a}}\overline{\psi\left(\frac{t-b}{a}\right)}dt \qquad (1)$$

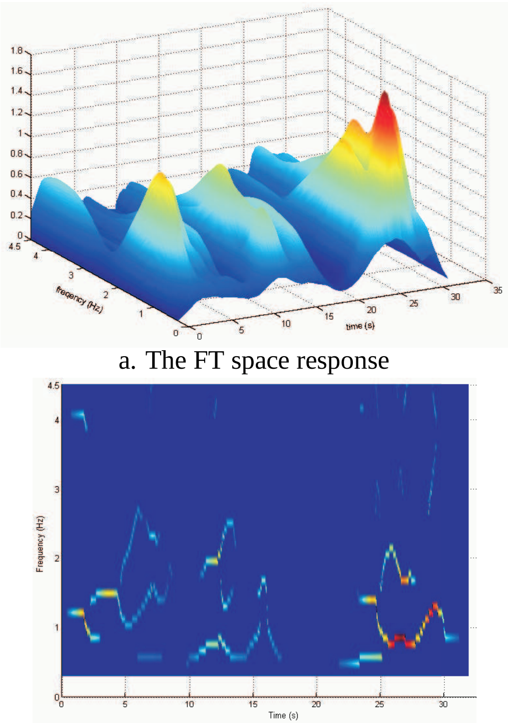
where $s(t)$ is the input signal and ψ is the wavelet function. a and b are the scale and position of the convolution window of ψ respectively[8]. By varying a, we can implement a filter bank of CWTs, producing a frequency-time space, FT. We experimented with various wavelet operators, and found the Morlet wavelet given by

$$\psi(t,a,b) = e^{(\frac{t-b}{a})^2/2}\cos\left(5\frac{t-b}{a}\right) \qquad (2)$$

most suitable for gesture oscillation extraction. Intuitively, this wavelet is constructed from a Gaussian operator convolved with a sinusoid. Hence, the Morlet wavelet has infinite duration (from the Gaussian) with most of the energy concentrated within the ‘hump’ of the Gaussian. Also this wavelet is centered, and the Gaussian performs an implicit smoothing operation over the input signal. $\psi(t,a,b)$ is essentially a band-pass filter centered around frequency $5/(2\pi a)$ with a bandwidth of $1/a$.

We set the values of a using a log scale: $[5/(2\pi \times 4.5)\ldots 5/(2\pi \times 0.3)]$. This produces a filter bank detecting center frequencies of the range: $[0.3\ldots 4.5]$Hz. At the high end, it is clear that the human hand is incapable of oscillating at 5 cycles per second (and even if it could, the camera could not pick up the motion). At the low end, 0.3 Hz is lower than a third of a cycle per second – beyond which gestural oscillations do not typically occur.

Figure 1.a shows the FT response of the y (up-down) oscillations of the right hand (RH) of a subject in a discourse dataset. Intersecting the FT surface with any constant time plane t_c produces the frequency response function $f_{t_c}(a)$ where a is the wavelet scale. Applying a smoothing filter to this signal will produce a series of scale peaks corresponding to dominant frequencies at t_c. These peaks trace out ridges through time. These maximal geodesic in FT space correspond to the dominant frequencies of the hand movement. To enhance the temporal continuity of these ridges in FT space, we first apply a cubic interpolating spline surface to the FT data before detecting the ridges by non-maximum suppression [9]. This produces a set of ridge traces as shown in figure 1.b. Notice that two different frequencies may be present simultaneously (as would happen if one were to move an arm up and down while ‘drawing’ smaller circles in the air). Notice, also, that a particular oscillation ridge may change in frequency. The algorithm does not require an oscillatory ridge behavior to be of a single frequency band – only that the ridge track a contiguous maximal geodesic in FT space. This matches the intuition that in oscillatory gestures, the hand is of finite mass and has to accelerate and decelerate through the gesture.



a. The FT space response

b. Frequency ridges

Fig. 1. Oscillation extraction of the subject’s RH y motion

3. EXPERIMENTAL SETUP

We performed our analysis on a dataset comprising 961 frame (32 sec) video of a subject describing her living space to an interlocutor (both seated) [3]. The data was captured with a monocular camera from an oblique frontal view.

We apply our parallel Vector Coherence Mapping (VCM) algorithm to obtain precise position traces both hands [10]. VCM tracks a large number of vectors (typically 600 to 2000 per frame) and integrates the fields. This averaging effect gives a smooth motion field that is temporally accurate (i.e. no oversmoothing across frames to degrade temporal resolution). Tracking errors are fixed manually.

We perform a detailed linguistic text transcription of the discourse that includes the presence of breath and other pauses, disfluencies and interactions between the speakers. We employ the Grosz ‘purpose hierarchy’ method [11] to obtain a discourse segmentation, and an alternate syntax-based analysis to find sentence breaks. The speech transcript is aligned with the audio signal using the Entropic’s word/syllable aligner. The output of the aligner is manually checked and edited using the Praat phonetics analysis tool [12] to ensure accurate time tags. This process yields a time-aligned set of traces of the hand motion with holds, head orientations and precise locations of the start and end points of every speech syllable and pause. The time base of the entire dataset is also aligned to the experiment video.

We extract the hand motions in the x and y dimension of the camera image plane. We label the respective hand-dimension pairs RH-x, RH-y, LH-x, and LH-y. Applying

Osc	#	Beg t	Dur.	Cyc.	$f_{R\max}$	Transcript
LH-x	1	6.57	0.90	2.29	2.54	… when you enter the hou[se]
	2	6.87	0.33	1.36	4.09	when you ent[er]
	3	9.61	1.03	2.29	2.21	… open the …
	4	12.65	0.93	3.82	4.09	um …
	5	15.11	2.41	4.17	1.73	… there's the the front staircase runs
	6	16.12	2.53	2.09	0.83	the front staircase runs right up there
	7	18.76	2.07	2.38	1.15	on on your left so you can go straight upst[airs]
	8	21.55	0.77	1.51	1.97	second floor from there
LH-y	1	16.14	2.11	1.92	0.91	the front staircase runs right up there
	2	18.74	2.12	2.44	1.15	on on your left so you can go straight ups[tairs]
RH-x	1	1.53	3.00	4.68	1.56	… so you're in the kitchen then there's a … the back stair oh I
	2	6.37	1.23	2.63	2.13	… when you enter the house
	3	9.51	1.24	2.23	1.80	… open the ...
	4	23.43	0.67	2.69	4.01	… but if you co[me]
	5	23.96	1.02	3.25	3.19	you come around through the kitchen
	6	24.72	0.26	1.06	4.09	kitchen
	7	27.16	2.67	1.76	0.66	winds around like this … and puts you up on the se[cond floor]
	8	28.16	1.67	3.15	1.89	[thi]s … and puts you up on the se[cond]
RH-y	1	1.33	0.79	0.97	1.23	… so you're in
	2	3.21	0.92	1.36	1.48	... the back sta[ir]
	3	25.49	2.20	4.69	2.13	[bac]k there's the back staircase that winds ar[ound]
	4	25.3	2.7	2.2	0.83	the back there's the back staircase that winds around like
	5	28	1.9	2.4	1.24	this ... and puts you up on the se[cond]

Fig. 2. Oscillation table

our wavelet ridge algorithm, we obtain oscillation ridge traces for the corresponding trace signals.

4. RESULTS AND DISCUSSION

Figure 2 tabulates the FT ridges. Each ridge is assigned a number. The table shows the beginning time, duration, and estimated number of oscillation cycles, the frequency at the point of highest wavelet response, $f_{R\max}$ (peak point of the ridge in FT space), and the words spoken by the subject within the ridge duration. The number of cycles is estimated from $f_{R\max}$ and the ridge duration. To show the temporal organization of these ridges with respect to the accompanying discourse, we 'scored' the ridge durations with the experiment transcript in figure 3.

The ridge durations are marked under each line of text in the order LH-x, LH-y, RH-x, RH-y, descending. When there are two simultaneous ridges in any FT space, they are scored one above the other (we did not detect any instance of more than two simultaneous ridges). For example, LH-x_5 and LH-x_6 begin simultaneously on the second line of the transcript, cotemporally with the words "…there's the the front …". By cross-referencing figure 3 with figure 2 we also know that $f_{R\max}$ for LH-x_5 and LH-x_6 are 1.73Hz and 0.83 Hz respectively.

The discourse piece begins with the subject describing the back of her house (kitchen and back staircase). The subject consistently situates this using her left hand [3]. The entire initial back of house is captured in the oscillatory segments RH-x_1 (1.56Hz). There were no perceptible oscillatory gestures in this segment. What we are measuring is the underlying frequency of her cadence (there were two phrases of approximately 3.7 sec duration during this period leading to the two motion cycles). This frequency ridge captures this entire back of house fragment right up to the point where she aborted the description with a rapid withdrawal of the RH at the beginning of the editing phrase: "Oh I forgot to say". RH-y_1 (1.23 Hz) lasted only one cycle and was basically the preparation for her first gesture and the motion of the gesture. RH-y_2 (1.48 Hz) was the twirling motion of the iconic gesture representing the spiral back staircase (we know that the staircase was spiral only from the gesture).

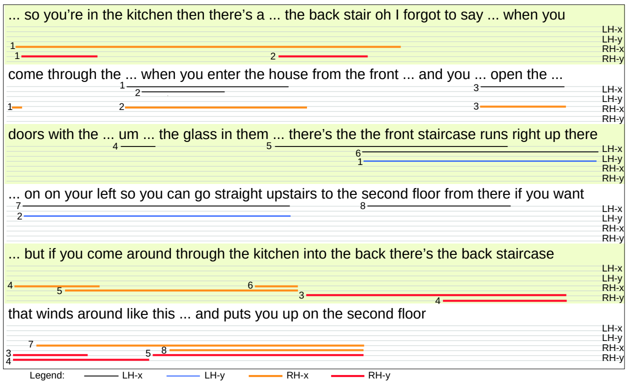


Fig. 3. Oscillation-scored Transcript

In the second phase of the discourse, the subject proceeds to the front of the house and the front doors. She consistently uses two-handed gestures to represent this discourse segment. This segment is clearly set out in the corresponding LH-x_1 (2.54 Hz) and RH-x_2 (2.13 Hz) symmetric 2H oscillatory gesture where she held both hands with horizontally with palms facing her torso at chest level (she was preparing to perform a gesture pantomiming "throwing open the double front doors"). The oscillatory gesture was a 'beat-like' gesture with hands thus held, possibly to indicate the facade of the house. This twice-repeated motion was properly captured. LH-x_2 (4.09 Hz) is a short high frequency oscillation we believe to be a noise ridge. The next strong symmetric oscillatory gesture, LH-x_3 (2.21 Hz) and RH-x_3 (1.8 Hz), captures the throwing open of the front doors (we know these were double doors again only from the gesticular imagery, it was never said). She performed the gesture with strong effort and the oscillation observed involved both the gesture and the 'overshoot-and-rebound' motion typical at the end of effortful gestural strokes.

In LH-x_4, the subject held both hands, palms out, in front of her in a 2H oscillatory action as though 'feeling' the glass doors. This is typical word search behavior, corresponding to the utterance "doors with the …<um> … the glass in them". There was a FT ridge of almost the exact shape in the RH-x, but the ridge peaks were weak, and we did not include it in the dataset. Hence, we deem the equivalent RH-x oscillation to be missed by the algorithm.

After describing the front of the house, the subject proceeded to describe the front staircase. In the discourse, she consistently used her LH gestures when describing this staircase. LH-x_5 – LH-x_8 and LH-y_1 and LH-y_2 pertain to this discourse segment. The subject performed two iterations of the same oscillatory gesture (2 motion cycles per gesture). The gesture indicated ascending straight staircases, with the hand rising and moving forward, with palm facing down and outward at approximately 45°. LH-x_6 (0.83 Hz) and LH-y_1 (0.91 Hz) captured the first gesture and LH-x_7 (1.15 Hz) and LH-y_2 (1.15 Hz) captured the second correctly. It is likely that the mental imagery contrast in this discourse was between the straight front staircase and the spiral back staircase. The final LH ridge, LH-x_8 (1.97 Hz), captures the final LH oscillatory gesture coincident with the words “second floor from there”. After two cycles of the front staircase gesture (at the end of LH-x_7 and LH-y_2), the subject held her LH at the highest point of the stroke, indicating the ‘upstairs’. As she uttered “second floor from there”, she oscillated her hand up and down at the wrist, palm down, as though patting the ‘upstairs floor’. LH-x_5 was a ridge that was barely strong enough to be counted. We believe that it was a noise ridge brought about by the strong neighboring ridges LH-x_6 and LH-x_7.

The subject returns to the back staircase and RH gestures in the final segment of the discourse. This segment begins with three relatively high frequency oscillations. In RH-x_4 (4.01 Hz), the subject raises her RH in a small twirling action. This is followed by RH-x_5 (3.19 Hz) where she makes a larger twirling gesture indicating ‘coming around to through the kitchen’. RH-x_6 (4.09 Hz) is a very quick hand-shape change (from an open down-facing palm in RH-x_5 to a typical ASL G-hand pointing pose for RH-y_3). This registered as a quick high frequency 1-cycle oscillation. RH-y_3 (2.13 Hz) is a return to the aborted spiral staircase gesture (upward twirling motion with hand in the pointing pose) of RH-y_2. RH-y_3 properly captures the two repetitions.

RH-y_4 (0.83 Hz) and RH-x_7 (0.66 Hz) both straddle two utterances. The ridge responses were strong, though there were no perceptible oscillatory gestures. We believe what we are capturing in each case is the ‘gestural cadence’ across two utterances. RH-x_8 (1.89 Hz) and RH-y_5 (1.24 Hz) detected the transition between the end of the spiral staircase gestures and the oscillatory beat that take the description to the second floor.

5. CONCLUSION

We have presented our wavelet-based extraction of oscillatory gestures. Our algorithm extracts the geodesic ridges of the FT space produced by a filter bank. We demonstrated the efficacy of the system in extracting natural human oscillatory gestures. We also presented a micro analysis of experimental discourse video to show the potential of such cross-modal analysis in the recovery of discourse structure.

6. REFERENCES

[1] D. McNeill, *Hand and Mind: What Gestures Reveal about thought*, U. Chicago Press, Chicago, 1992.

[2] D. McNeill, “Growth points, catchments, and contexts,” *Cognitive Studies: Bullet. Japanese Cog. Sci. Soc.*, vol. 7, no. 1, 2000.

[3] F. Quek, D. McNeill, R. Ansari, X. Ma, R. Bryll, S. Duncan, and K-E. McCullough, “Multimodal human discourse: Gesture and speech,” *ToCHI*, In Press.

[4] F. Quek, “The catchment feature model: A device for multimodal fusion and a bridge between signal and sense,” In Review: *EURASIP JASP*.

[5] G. Fauconnier, *Mental Spaces: A spects of Meaning Construction in Natural Language*, MIT Press, 1985.

[6] F. Quek, D. McNeill, R. Bryll, and M Harper, “Gestural spatialization in natural discourse segmentation,” in *7th Int Conf on Spoken Language Proc.*, 2002, pp. 189–192.

[7] L.S. Vygotsky, “Thinking and speaking,” in *The Collected Works of L.S. Vygotsky, Vol.1, Problems in General Psychology (N. Minick Trans.)*, R.W. Rieber and A.S. Carton, Eds., pp. 39–285. Pleunm, 1987.

[8] R.M. Rao and A.S. Bopardikar, *Wavelet Transforms: Introduction to Theory and Applications*, Addison-Wesley, 1998.

[9] J. Canny, “A computational approach to edge detection,” *PAMI*, vol. 8, no. 6, pp. 679–698, 1986.

[10] F. Quek, X. Ma, and R. Bryll, “A parallel algorithm for dynamic gesture tracking,” in *ICCV’99 Wksp on RATFG-RTS.*, 1999, pp. 119–126.

[11] C.H. Nakatani, B.J. Grosz, D.D. Ahn, and J. Hirschberg, “Instructions for annotating discourses,” Tech. Rep. TR-21-95, Ctr for Res. in Comp. Tech., Harvard U., MA, 1995.

[12] P. Boersma and D. Weenik, “Praat, a system for doing phonetics by computer,” Tech. Rep. Report 132, Institute of Phonetic Sciences of the University of Amsterdam, 1996.