Using Optical Flow and Electromagnetic Articulography in Multimodal Speech Research

Samantha Gordon Danner[1], Louis Goldstein[1], Eric Vatikiotis-Bateson[2], Robert Fuhrman[2], Adriano Vilela Barbosa[3]

[1]University of Southern California, USA
[2]University of British Columbia, Canada
[3]Federal University of Minas Gerais, Brazil

sfgordon@usc.edu

Abstract

This paper describes a technique for studying the coordination of different modalities in multimodal speech using audio, video, and kinematics data. We used an electromagnetic articulograph (EMA) and audio/video recording equipment to collect data. The data is processed using EMA data analysis software and software called FlowAnalyzer, which employs the computer vision technique *optical flow* to estimate the velocity of video-recorded movement. Motion coordination between speech and manual gesture at various temporal offsets can then be analyzed using correlation map analysis and other methods. The approach outlined below represents an accessible way to study various aspects of multimodal communication.

Index Terms: Spontaneous speech, gesture, electromagnetic articulography, optical flow

1. Introduction

There is a long tradition of research on speech and co-speech gesture (henceforth *multimodal speech*) using video and audio recordings to measure coordination of acoustic features of speech with visually-observed gestures of, e.g., the hands, arms, and head [1]. These studies often report compelling results that suggest the presence of high-level coordination between speech acoustic events and body movements [2].

One potential shortcoming in many of these studies is that different kinds of measurements and temporal landmarks are used in each methodology [2]. Additionally, though there are highly developed coding methodologies for co-speech gesture research [3, 4], important details of the coding strategy – such as phases of gestural movements – can be interpreted variably, dependent on the experiment, researcher, or theory. Another critical issue is the naturally fluctuating coordination found in biological systems [5]: as exactly synchronic movements rarely occur in nature, findings of precise coordination are quite rare. Researchers may therefore wish to address the likelihood of movement correlation at temporal offsets.

The methods described here make it possible to investigate and quantify multimodal speech in novel ways. A goal of this work is to make the process of collecting and analyzing multimodal speech data more accessible to researchers. To address the concern of landmark-based measurements, we describe a method that allows for time-varying comparison of speech articulator and bodily gesture movement velocities. These velocity measurements provide an alternative to problematic coding strategies: velocity peaks – indicative of movement amplitude – in different movement signals can be compared without making assumptions about movement initiation and completion or other arbitrary landmarks in the auditory or visual domains of analysis. We also describe how correlation map analysis [5] and other tools can be used to analyze naturally fluctuating coordination at temporal offsets.

The rest of the content of this paper is organized as follows: Section 2 describes some findings and open questions in multimodal speech research. Section 3 describes our methods for studying multimodal speech coordination. Section 4 outlines some research findings and future directions for the methodology under discussion here.

2. Multimodal Speech

In experimental speech research, acoustic studies dominate the field. Acoustic data is easy to collect and analyze, but the biggest limitation is that only one modality of speech can be considered. As video recording and motion capture technology has become widely available, these tools have been integrated into speech research, but data collection and analysis of multimodal data is more complex than the analysis of acoustics alone. Studying auditory, visual *and* kinematic modes of speech together can provide a holistic view of the production and perception of speech, as there has long been an awareness that visual information [6], speech rate and amplitude modulation [7], facial expression [8] and body posture and gestures [4] are all critically important to speech perception and production.

In the domain of speech acoustics and speech articulator kinematics, some findings about temporal coordination and stereotypical gestural units [9] have become recognized as standard by researchers, but the same cannot be said for manual gesture. One issue is a lack of agreement in the research community about gesture classification schemes [10]. These schemes may disagree about how initiation and completion of gestures are defined, hand-specificity (as manual gestures may involve the use of the right hand, left hand, or both), how to handle 'combination' gestures (composed of two or more *types* of gesture), etc. A few studies that have taken on the task of researching multimodal speech in laboratory settings are described below.

2.1. Multimodal speech in the laboratory

Coordination of speech and manual movement has been observed in certain directed tasks in laboratory settings. Because of the inherent difficulties in the segmenting of complex time-varying co-speech gestures, many researchers have studied specific types of manual gesture in isolation, such as pointing or *deictic* gestures [4]; rhythmic finger

tapping is another prevalent gestural paradigm that researchers use to stand in for more complicated gestures. Participants in such studies are typically asked to start movement of the hand or finger from, and return to, a base position to make analysis of gesture initiation and completion less ambiguous. These studies have found, across a variety of measurement paradigms, that deictic gestures and finger tapping motions are significantly correlated with positions of prosodic prominence in the speech signal [11, 12]. Such results are highly encouraging, but they suggest the need for studying a wider variety of gestures in more naturalistic speech contexts.

2.2. Open questions in multimodal speech research

Manual gesture has many more degrees of freedom than does speech articulation, and its' use is generally arbitrary. The use of manual gesture also appears to be idiosyncratic and/or culturally influenced [13]. Some researchers have questioned the underlying purpose of manual gesture, ascribing its use to, e.g., lexical retrieval rather than communication, given that speech-accompanying gesture on its own may be less communicative than speech itself [14].

The speech and gesturing tasks sometimes used in multimodal speech experimentation also raise concerns about ecological validity, especially with regard to spontaneous speech and interpersonal communication. To generate testable hypotheses regarding the coordination of speech and manual gesture, some studies investigate coordination through higher-level phenomena such as speech prosody [11]. Findings in this line of research are promising, but it is probably the case that prosody is not the only process controlling the deployment and coordination of manual gesture (respiration has been shown to play a role [15], for example). The causal relationship between speech prosody and gestural coordination is also not well understood. It is thus necessary to find a way to research coordination in a variety of communication tasks and environments.

3. Methods for studying multimodal speech coordination

The research presented here is proof of concept for the simultaneous recording of EMA kinematics data, and video data transformed into motion data with the use of optical flow software [12, 13]. Optical flow (OF) is a computer vision technique that tracks changes in pixel intensity across frames in a video sequence. OF algorithms track the magnitude and direction of these intensity changes across video frames, thus allowing for the recovery of velocity and direction of motion. FlowAnalyzer, the software that is used in this analysis, allows for post-hoc selection of regions of interest within a video prior to processing data. The software creates signals that can be additionally processed using a set of MATLAB tools [16] in subsequent steps of analysis. Minimally, a video camera with a microphone and a tripod is sufficient equipment to collect movement data with a reasonable degree of accuracy (compared with marker-based tracking [17]), thus allowing for easily portable experiment setups and research in the field. OF measurements also have the benefit of being non-invasive, so they can be used to easily track motion in regions that cannot be tracked by EMA, either due to the limited size of the equipment's magnetic field, or due to limitations on the number of sensors that can be tracked simultaneously. Body movement regions of interest measured with OF need not be predetermined, unlike with flesh point measures. The combination of EMA and video-derived OF measures allows for the direct comparison of speech articulator motion and motion of manual/bodily gesture. Previously, obtaining measures to make this comparison has been much more time-consuming and limited in scope.

3.1. Data Acquisition

The research described here investigates whether properties of gesture and coordination of speech and speech-accompanying manual gesture are task-dependent. This question was motivated by the observation that some speech situations may necessitate the use of manual gesture to a greater degree than others [18], and that different speech tasks may demand different kinds of speech-gesture coordination. The experiment is briefly described below.

3.1.1. Materials

EMA data was collected using an electromagnetic articulographer (WAVE, Northern Digital), sampled at 400 Hz. Audio, sampled at 44.1kHz, was collected concurrently using a microphone synchronized with the EMA equipment. EMA sensors were placed on the tongue body, tongue tip, lower incisor (jaw), and the upper and lower lip. Additional reference sensors were placed on the left and right mastoid processes and the upper incisor. A GoPro Hero4 video camera on a tripod positioned approximately 2' from the participant, with the speaker's head, shoulders, arms and torso in the field of view, was used to capture video at 29.97fps and audio at 48kHz.

3.1.2. Method

Participants (*N*=3) were seated comfortably in an armless chair (so as not to impede movement of the arms and hands), positioned in front of a computer monitor and next to the EMA magnetic field generator. EMA sensors were attached to the participant and the EMA equipment was calibrated. The experiment was presented in two randomized blocks. In the first block (the *demo task*), the participant was asked to demonstrate to the experimenter how they would perform an action such as eating a banana, or opening an umbrella (henceforth, these actions will be referred to as *themes*). In the second block (the *response task*), participants answered preference questions asked by the experimenter regarding each theme (e.g., "Do you prefer to eat bananas sliced or whole, and why?" or "Do you prefer umbrellas with a button closure or a Velcro closure, and why?"). Participants had a clear view of both the experimenter and the monitor. Instructions for each trial, accompanied by a black and white line drawing of the trial's theme, were presented on the monitor. The use of manual gesture was never explicitly mentioned, although the experimenter demonstrated a practice trial for each block with the use of speech and manual gesture.

3.2. Data Processing

Raw EMA data files were processed in kinematic analysis software (Mview, Haskins Laboratories), from which tangential velocities for each EMA sensor were calculated. A Praat Textgrid with interval tiers for acoustic duration, words, pauses and comments was created and annotated for each trial[19]. Praat was also used to cross-correlate audio from different sources for signal alignment.

Videos were segmented into trial-length segments and converted to .mov format at 30fps, 1280x720pixel resolution,

with 44.1kHz mono channel audio (note that video need not be high resolution to work with OF). Video files were then processed in OF analysis software (e-mail the corresponding author for information on how to obtain the free FlowAnalyzer software), which creates a number of signal files that can be manipulated in MATLAB using the Audiovisual Speech Processing (AVSP) toolbox. The AVSP toolbox was used to create MATLAB structures containing the signal information obtained from the OF analysis completed in the previous step.

3.3. Data Analysis Tools

When EMA tangential velocity, acoustic segmentation and OF signals have been processed, time-varying correspondence can be analyzed using correlation map analysis (CMA) [5]. The signals generated from the AVSP toolbox and Mview can be manipulated in MATLAB or other programs to obtain measurements of interest. For example, we used MATLAB to analyze velocity time series for each movement signal (EMA sensors and OF regions of interest) in each trial. This velocity information was used to create measures of peak velocity and path length, and to measure correlations and causality in right hand and jaw velocity signals. Some findings from this analysis are described below, in section 4.

4. Findings and Future Directions

Comparison of peak velocities in the *demo* task and the *response* task (see 3.1.2) suggest that use of manual gesture may indeed be task-dependent. The peak velocities of articulator movements (Right Hand/RH and Jaw are discussed here) represent a way of measuring the average amplitude of movements without the need to predefine movement anchor points. We found that Jaw peak velocities did not differ significantly as a function of the type of speech task participants were engaged in ($t(2)=0.861$, $p=0.4798$), but average RH peak velocities were significantly greater in the *demo* task than the *response* task ($t(4.92)=2.599$, $p=0.0491$); see Figure 1. This result shows that the effect of task has a greater impact on manual gesture than it does on speech articulatory gestures, and the result was obtained without the need for a priori assumptions about manual gesture typology.

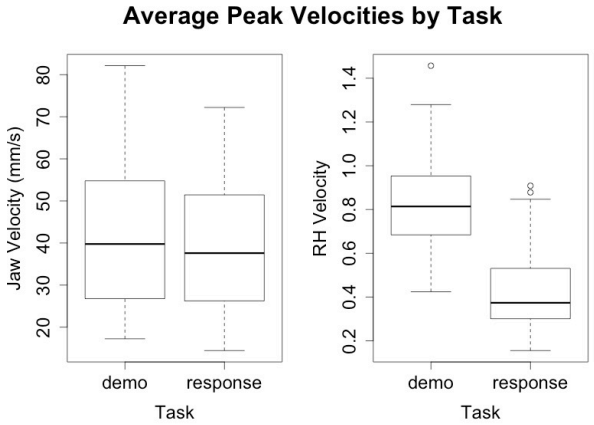

Figure 1: *Average peak velocities of movement signals in two tasks (all participants)*

We were also interested in assessing the causal relationship between two movement signals, in this case, the jaw and the right hand. We have applied the Granger test to measure causality, where a signal *X* Granger-causes a signal *Y* "if an auto-regressive model for *Y* in terms of past values of both *X* and *Y* is statistically significantly more accurate than that based just on the past values of *Y*" [20]. We applied this test to the right hand (RH) and jaw signals in each experiment condition for one theme in one speaker's data. In this small sample, the only significant result of Granger causality (see Table 1) was found in the Response condition. There, the RH signal was found to Granger-cause the Jaw signal, indicating that the instantaneous velocities of the RH signal were predictive of instantaneous velocities in the Jaw signal, at a short lag (0.115s) with respect to the RH signal. This indicates an especially close relationship between movements of the right hand and the jaw in the *response* condition, but only in one of the two possible orders. In the same condition, Jaw movements were *not* predictive of RH movements at the same short lag. The causality findings suggests that (Granger) causal relationships between speech articulator and manual movements are task-dependent, and that speech-accompanying gesture in the *response* task is more closely coordinated with speech than is gesture in the *demo* task.

Table 1: *Granger Causality tests on RH & Jaw signals in speaker M1's "fold laundry" theme*

	RH → JAW	**JAW → RH**
Demo	$F(16, 4437)=0.5681$, $p=0.9095$; lag=0.08s	$F(16,4437)=0.2136$, $p=0.9996$; lag=0.08s
Response	$F(23,13095)=2.0134$, $p=0.0028$*; lag=0.115s	$F(23,13095)= 0.2974$, $p=0.9995$; lag=0.115s

Above, we've described methods for easily obtaining measurements and comparisons of selected speech articulators and movements of the head and hands. Nearly any movement region of interest can be measured with the OF technique described here. Another benefit of this technique is that it allows for experimental designs that make use of realistic communicative events. We have also show that spoken language and manual speech-accompanying gesture each have a directly comparable kinematic component, and the relationship between the kinematic components may be used to investigate hypotheses about coordination of speech and gesture. In the case described in section 3, we have used video recording of manual and head gesture and EMA to record speech articulatory kinematics from specific points on the vocal tract. There are many other areas of study in multimodal speech research, a few of which are outlined below.

4.1. Additional scales of comparison

The OF techniques described here could be used on finely detailed movements of the face (the perioral region and the eyes and eyebrows are known to be particularly informative [8], [17]). For example, one might want to corroborate EMA recordings of lip or jaw kinematics with video recordings of the perioral region to ensure reasonable correlation of the two motion signals. Smaller regions of interest around the eyes may also be considered: As discussed in [17], potentially coordinated movement behaviors such as blinking can be captured with OF, but not with other marker-based systems.

Motion capture and OF can similarly be used for studying coordination at even larger scales. Prior research has found, for example, a relationship between postural control and vocal effort in speech [15]. This research hints at the expansive coordination of systems within the human body. Findings in this area could also help reveal what makes certain

gesturing body parts like the arms, hands and head 'special' in multimodal speech communication.

4.2. Studying multimodal speech cross-linguistically

Many advances in the field of multimodal speech/communication have been driven by the need to investigate sign languages [21] in a principled fashion. Because (visually-observed) movements are a primary modality in sign language, the need for studying coordinated movement behavior in sign language is obvious.

Another possibility to investigate is cross-linguistic coordination and/or timing differences in various speech modalities. As EMA research has uncovered that some of the detailed timing and organization of speech articulator gestures is language-specific [22], it is reasonable to expect that comparable cross-linguistic differences exist in the timing of manual gestures with respect to speech articulator gestures and/or other bodily gestures.

4.3. Conversational Interaction

A final scenario to consider is the interaction of two or more speakers engaged in a shared speech task. The possibility of using dual EMA systems to simultaneously record vocal tract kinematics from two interacting participants has been validated [23], and the ability to video record and compute optical flow on regions of interest in conversational interactions has also been demonstrated [24]. These researchers and many others have found *entrainment* between speakers in conversation occurring at various levels. Entrainment generally refers to the rhythmic alignment that occurs between motor subsystems; the ability of the subsystems to become entrained is thought to be an indication of the presence of a higher-level dynamical system governing these subsystems. The methods described here could be applied to entrainment phenomena in multimodal speech.

5. Summary

As workflows like the one presented here become easier to implement, and as motion capture and video recording equipment becomes more affordable and accessible for researchers, we expect that the analysis of movement and kinematic data in multimodal speech research will become ubiquitous. The goal of this work is to provide some suggestions and resources for collecting and analyzing multimodal speech data, and to that end we have described research scenarios that could make use of this type of data.

6. Acknowledgments

This work was supported by NSERC and SSHRC grants to Eric Vatikiotis-Bateson and an NIH grant to Dani Byrd. The EMA data described in this paper was collected in collaboration with Sungbok Lee.

7. References

[1] S. Shattuck-Hufnagel, P. L. Ren, and E. Tauscher, "Are torso movements during speech timed with intonational phrases?," in *Proeedings of Speech Prosody 2010*, 2010, pp. 2–5.

[2] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol. 57, pp. 209–232, 2014.

[3] S. Duncan, "Annotative Practice (Under Perpetual Revision)," in *Gesture & Thought*, 2005.

[4] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[5] A. Barbosa, R.-M. Déchaine, E. Vatikiotis-Bateson, and H. Yehia, "Quantifying time-varying coordination of multimodal speech signals using correlation map analysis," *J. Acoust. Soc. Am.*, vol. 131, no. 3, p. 2162, 2012.

[6] H. McGurk and J. Macdonald, "Hearing lips and seeing voices.," *Nature*, vol. 264, pp. 691–811, 1976.

[7] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *J. Acoust. Soc. Am.*, vol. 52, no. 4, pp. 1238–1250, 1972.

[8] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," *... Semin. Speech Prod. (ISSP 2006)*, no. 2003, pp. 549–556, 2006.

[9] L. Goldstein and M. Pouplier, "The Temporal Organization of Speech," in *The Oxford Handbook of Language Production*, M. Goldrick, V. S. Ferreira, and M. Miozzo, Eds. Oxford University Press, 2014, pp. 1–21.

[10] J.-P. de Ruiter, "Gesture and speech Production," 1998.

[11] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Temporal coupling between speech and manual motor actions.," *9th International Seminar on Speech Production*. 2011.

[12] J. Krivokapic, M. K. Tiede, and M. E. Tyrone, "A Kinematic Analysis of Prosodic Structure in Speech and Manual Gestures," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[13] S. Kita, "Cross-cultural variation of speech-accompanying gesture: A review," *Lang. Cogn. Process.*, vol. 24(2), no. December 2014, pp. 145–167, 2009.

[14] R. M. Krauss and U. Hadar, "The Role of Speech-Related Arm/Hand Gestures in Word Retrieval," *Gesture, speech, sign*, pp. 93–116, 1999.

[15] R. Fuhrman, "Vocal effort and within-speaker coordination in speech production: effects on postural control," University of British Columbia, 2014.

[16] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "MATLAB Toolbox for Audiovisual Speech Processing," in *AVSP 2007*, 2007.

[17] A. V. Barbosa, H. C. Yehia, and E. Vatikiotis-Bateson, "Linguistically Valid Movement Behavior Measured Non-Invasively," *Audit. Vis. Speech Process.*, pp. 173–177, 2008.

[18] R. B. Church, S. Kelly, and D. Holcombe, "Temporal synchrony between speech, action and gesture during language production," *Lang. Cogn. Neurosci.*, vol. 29, no. 3, pp. 345–354, Nov. 2013.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." 2016.

[20] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '07*, p. 66, 2007.

[21] O. Crasborn, H. Sloetjes, E. Auer, and P. Wittenburg, "Combining video and numeric data in the analysis of sign languages within the ELAN annotation software," *Proc. Lr. 2006 Work. Represent. Process. sign Lang.*, pp. 82–87, 2006.

[22] L. Goldstein, I. Chitoran, and E. Selkirk, "Syllable Structure as Coupled Oscillator Modes: Evidence from Georgian vs. Tashlhiyt Berber," *Proc. XVI Int. Congr. Phonetic Sci.*, pp. 241–244, 2007.

[23] E. Vatikiotis-Bateson, A. V. Barbosa, and C. T. Best, "Articulatory coordination of two vocal tracts," *J. Phon.*, vol. 44, pp. 167–181, May 2014.

[24] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. G. Munhall, "Movement coordination during conversation," *PLoS One*, vol. 9, no. 8, pp. 1–10, 2014.