



Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Quantitative analysis of multimodal speech data

Samantha Gordon Danner^{a,*}, Adriano Vilela Barbosa^b, Louis Goldstein^a

^a University of Southern California, Los Angeles, CA 90089, USA

^b Federal University of Minas Gerais, Belo Horizonte, MG CEP 31270-901, Brazil



ARTICLE INFO

Article history:

Received 2 October 2017

Received in revised form 10 September 2018

Accepted 26 September 2018

Keywords:

Multimodal speech

Bodily gesture

FlowAnalyzer

Correlation Map Analysis

Time-varying coordination

Communicative context

ABSTRACT

This study presents techniques for quantitatively analyzing coordination and kinematics in multimodal speech using video, audio and electromagnetic articulography (EMA) data. Multimodal speech research has flourished due to recent improvements in technology, yet gesture detection/annotation strategies vary widely, leading to difficulty in generalizing across studies and in advancing this field of research. We describe how FlowAnalyzer software can be used to extract kinematic signals from basic video recordings; and we apply a technique, derived from speech kinematic research, to detect bodily gestures in these kinematic signals. We investigate whether kinematic characteristics of multimodal speech differ dependent on communicative context, and we find that these contexts can be distinguished quantitatively, suggesting a way to improve and standardize existing gesture identification/annotation strategy. We also discuss a method, Correlation Map Analysis (CMA), for quantifying the relationship between speech and bodily gesture kinematics over time. We describe potential applications of CMA to multimodal speech research, such as describing characteristics of speech-gesture coordination in different communicative contexts. The use of the techniques presented here can improve and advance multimodal speech and gesture research by applying quantitative methods in the detection and description of multimodal speech.

© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The study of linguistics has long been centered on aspects of spoken communication, such as the sound, structure, and meaning of words and utterances. Though these topics are the foundation on which the study of linguistics has been built, speech alone does not always communicate the entirety of a message. There are many other elements that speakers use to communicate effectively, such as manual gestures, facial expressions, body posture, etc. (Busso & Narayanan, 2006; Garrod & Pickering, 2004; McNeill, 1992; Shattuck-Hufnagel, Ren, & Tauscher, 2010; Yehia, Rubin, & Vatikiotis-Bateson, 1998). Evidence from multimodal speech research suggests that speech-accompanying bodily movements share important “neurological and biomechanical linkages” with speech articulator movements (Barbosa, Déchaine, Vatikiotis-Bateson, & Yehia, 2012), and that speakers use bodily movements for various communicative purposes, such as to express affiliated semantic concepts in the speech signal, or to aid in the prosodic structuring of communication (Ferré, 2010; Krahmer &

Swerts, 2007). It has also been shown that bodily movement may facilitate lexical access (Rauscher, Krauss, & Chen, 1996). And, while some non-speech communicative modalities – particularly manual gesture – have a long history of being studied qualitatively/observationally (Bolinger, 1968; Friesen, Ekman, & Wallbott, 1979; Kendon, 1970; McNeill, 1992), the quantitative study of non-speech communicative modalities in linguistics is still coming into maturity.

Linguists are increasingly integrating research on non-speech modalities into the study of fields like semantics, prosody, and information structure, using various strategies to describe and annotate bodily movements, or *gestures*, as they relate to speech (Enfield, Kita, & de Ruiter, 2007; Ferré, 2010; Krivokapić, Tiede, & Tyrone, 2017). While variation in annotation strategy is expected because there are so many different research questions in this field (Wagner, Malisz, & Kopp, 2014), there is a clear need to achieve some degree of standardization in the identification and description of speech-accompanying gesture, in order to generalize and theorize about the relationship between speech and bodily gesture across different studies and research questions. In this paper, we describe two software tools, FlowAnalyzer and Correlation Map Analysis (CMA), which together offer an accessible

* Corresponding author.

E-mail address: sgordondanner@gmail.com (S.G. Danner).

methodology for detecting and quantifying linguistically valid multimodal speech data. We describe a multimodal speech experiment utilizing FlowAnalyzer and CMA, including some exploratory analyses of the resulting data. With this description of quantitative techniques for studying communicative movement, we hope to show how feasible it is to integrate bodily gesture and multimodality research into broader linguistic research programs.

1.1. Collecting multimodal speech data

One potential barrier to entry for researchers interested in gesture and multimodal speech is the need to acquire audiovisual or kinematic data for analysis. FlowAnalyzer,¹ one of the tools we describe in this paper, can produce motion data from digital video recordings, even at low resolutions (Barbosa & Vatikiotis-Bateson, 2013; Barbosa, Yehia, & Vatikiotis-Bateson, 2008). In addition, FlowAnalyzer can be used with existing multimodal speech corpora (Busso et al., 2008; Shattuck-Hufnagel et al., 2010) or even YouTube videos for multimodal speech and gesture research. FlowAnalyzer runs on Linux, Mac, and Windows operating systems; in addition, an easy-to-use, standalone version of the software is available for Mac.

FlowAnalyzer uses optical flow, a computer vision technique (Horn & Schunck, 1981), to track motion in video. The optical flow algorithm compares consecutive frames of the video sequence and, for each pixel, calculates a displacement vector corresponding to the pixel intensity differences across the two frames. The displacement vectors are represented as Cartesian (horizontal, vertical) or polar (magnitude, direction) coordinates. The array of displacement vectors across all frames comprises the optical flow field (Barbosa et al., 2008). In addition to the basic implementation of optical flow, the FlowAnalyzer software also gives users the ability to select, *post-hoc*, multiple regions of interest and disinterest within a video (Barbosa & Vatikiotis-Bateson, 2013); it is thus possible to distinguish (or exclude) movement in multiple regions of a video simultaneously, as we have done in the research described here by tracking a speaker's head and hands in separate regions. FlowAnalyzer requires a stable background across video frames, so it should be used with video recorded from a stationary camera. The output optical flow signals have the same frame rate as the video recordings submitted to FlowAnalyzer, so researchers may wish to modify their video frame rate before submitting recordings to FlowAnalyzer, depending on their analysis goals. Other tools similar to FlowAnalyzer are available (Paxton & Dale, 2013; Westlund, D'Mello, & Olney, 2015), but we believe FlowAnalyzer's ease of use, flexibility, and region of (dis)interest selection capabilities make it a convenient general purpose tool for two-dimensional motion tracking.

Other affordable markerless motion tracking options include Microsoft Kinect, which provides depth mapping and pose estimation capabilities (Namboodiripad, Lenzen, Lepic, & Verhoef, 2016), and key point or joint estimators (Simon, Joo, Matthews, & Sheikh, 2017). Those who study speech articulation are likely already familiar with the use of electromagnetic articulo-

graph (EMA) systems and other point- or marker-based motion tracking tools. Such tools allow for the three-dimensional study of individual points on the articulators. Whatever motion tracking technique is used, kinematic information is a valuable resource in multimodal speech research because it allows for direct investigation of movement dynamics. Many phonetics researchers are employing motion-tracking data in their research (Krivokapić et al., 2017; Parrell, Goldstein, Lee, & Byrd, 2014; Rochet-Capellan, Laboissière, Galván, & Schwartz, 2008; Roustan & Dohen, 2010; Vatikiotis-Bateson, Barbosa, & Best, 2014; Yehia et al., 1998), and in the rest of this paper we demonstrate how kinematic data can be used together with other acoustic or video data for phonetic and linguistic investigations.

1.2. Identifying and annotating speech-accompanying gesture

Another difficulty of performing linguistic research on gesture is deciding how to identify and annotate bodily gesture from a continuous stream of movements. Annotation of bodily gesture requires numerous decisions, such as determining where one gesture ends and the next begins (Duncan, 2005), identifying the lexical affiliate of a bodily gesture in speech (Ferré, 2010), deciding whether a gesture exemplifies a proposed type (Hostetter, Alibali, & Kita, 2007), etc. The identification and annotation of bodily gesture is time-consuming and highly variable (Wagner et al., 2014), because gesture researchers have different aims and use a variety of signals to study multimodal speech (Westlund et al., 2015). Because not all speech-accompanying gestures are as conventionalized as spoken language is, gesture identification necessitates the use of qualitative judgments. Many researchers use multiple annotators and provide inter-rater reliability statistics to address such concerns (Busso et al., 2008; Friesen et al., 1979), but even with these safeguards, the range of gesture annotation techniques makes it difficult to directly compare results across studies.

Multimodal speech studies often identify and analyze a particular type of bodily gesture, such as *beat* (Krahmer & Swerts, 2007; Leonard & Cummins, 2011), *iconic/representational* (Beattie & Shovelton, 2000; Hostetter et al., 2007), or *deictic* gestures (Krivokapić et al., 2017; Rochet-Capellan et al., 2008). *Beats* are typically described as small, quick, “biphasic” movements made with formless hands; *deictics* are pointing movements, typically involving an extended index finger; and *iconic/representational* gestures are movements that depict some element of the accompanying speech (McNeill, 1992). McNeill's influential gesture typology (1992, pp. 78–80) uses *observational* details of manual gesture; as yet, there are few studies – but cf. Roustan & Dohen (2010) – that have considered quantitative or kinematic properties of gesture within the McNeill typology. We believe that applying quantitative methods will improve the annotation and description of speech-accompanying gesture.

To demonstrate how quantitative methods can improve multimodal speech research, we show that it is possible to detect the occurrence of individual bodily gestures and to assess quantitative properties of gesture as a function of communicative task. In research on the velocity profiles of speech articulator and limb movements, it has been shown that a simple

¹ FlowAnalyzer is available at www.cefa.org/FlowAnalyzer, along with some documentation.

‘gesture,’ i.e., lowering of the mandible or extension of the forearm, is associated with a velocity profile (Munhall, Ostry, & Parush, 1985; Ostry, Keller, & Parush, 1983), which typically has a single peak (instantaneous velocity maximum). This research has also demonstrated that the velocity peak of a movement is closely associated with movement amplitude (Munhall et al., 1985). We straightforwardly apply these findings to head and manual movements in the present research for the near-automatic detection of speech-accompanying gestures: here, an individual gesture is simply associated with a velocity peak. Although this strategy is simplistic – for example, what is perceived as a single gesture may have more than one velocity peak (Saltzman & Munhall, 1989) – the research we describe below suggests that this method is useful as a first approximation for detecting and measuring speech-accompanying gesture. We apply this strategy to study whether characteristics like gesture peak velocity magnitude or the likelihood of correlation between speech and manual gesture signals serve to distinguish multimodal speech produced in distinct communicative tasks in a quantifiable way.

1.3. The relation between speech and bodily gesture

Two prevalent topics in multimodal speech research are *when* and *why* manual gestures accompany speech. Some researchers have studied *anchor points* where speech and manual gesture temporally coordinate or attract one another (Leonard & Cummins, 2011; Wagner et al., 2014). Few researchers have reported evidence of anchor points at which speech and manual gestures consistently coordinate (Leonard & Cummins, 2011; Rusiewicz, Shaiman, Iverson, & Szuminsky, 2014). The scarcity of exact cross-modal co-occurrence may be explained with reference to the observation that biokinematic signals naturally fluctuate (Barbosa et al., 2012). Many researchers have thus taken a wider view of co-occurrence between speech and manual gesture (Leonard & Cummins, 2011; Loehr, 2007; Rusiewicz, 2011), noting a “gesture lead” effect, wherein manual gesture onsets often precede the onset of accompanying speech (Nobe, 2000; Rusiewicz et al., 2014). Similarly, a looser notion of cross-modal alignment has led to the discovery of cross-modal links occurring near speech events such as stressed or accented vowels, syllables, words or phrases (Leonard & Cummins, 2011; Yasinnik, Renwick, & Shattuck-Hufnagel, 2004). Rochet-Capellan et al. (2008) showed, for Brazilian Portuguese, that pointing gestures started before jaw opening gestures, at a delay dependent on stress placement in the speech. Leonard and Cummins (2011) found, for British English, that the movement onset of beat gestures occurred before the acoustic onset of the following stressed vowel in the speech signal. For American English, Nobe (2000) found that 23.9% of representational gestures had onsets in silent pauses preceding or between speech periods. There are several reports about which part(s) of a manual gesture co-occurs with speech: proposals include the maximum *spatial* extension of a gesture (the *apex*) or the time at which a change in direction of movement is observed (Leonard & Cummins, 2011; Yasinnik et al., 2004). These findings suggest that imprecise temporal coordination is a feature of the relationship between speech and bodily gesture (Leonard & Cummins, 2011; Rusiewicz

et al., 2014). In the present paper, we use Correlation Map Analysis (CMA) to investigate the correlation between speech and gesture at a range of temporal delays, to understand the nature of the time-varying relationship between speech and bodily gesture.

Many gesture researchers have built solid cases in support of semantics, prosody, and speech planning as possible functions of or explanations for *why* bodily gesture cooccurs with speech. There are strong arguments for the semantic connection between speech and manual gesture (de Ruiter, 1998; Ozyurek, 2014); some gestures clearly visually depict aspects of the concurrent speech. Other researchers note the association between some manual gestures and prosodic prominences in speech (Krahmer & Swerts, 2007; Krivokapić et al., 2017), suggesting that manual gestures represent a visual rhythm, or that bodily gestures might serve as ‘visual prosody’ (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Still other researchers demonstrate that speech-accompanying gestures are likely to facilitate lexical access or serve speech planning processes (Butterworth & Beattie, 1978; Rauscher et al., 1996). Bodily gestures likely serve all of these functions and more (Wagner et al., 2014).

In the experiment described below, we study the *quantitative* and *coordinative characteristics* of speech and speech-accompanying gesture kinematics in two qualitatively different communicative tasks, and we aim to use characteristics of multimodal speech to distinguish the communicative tasks. We find systematic differences in manual movement characteristics and in cross-modal coordination characteristics across communicative tasks, suggesting that it is possible to look more systematically at the relationship between quantitative characteristics of multimodal speech and existing gesture annotation techniques. We believe this to be a reason to conduct further research on the kinematics of multimodal speech, with the goal of relating quantitative aspects of multimodal speech behavior to bodily gesture as it has traditionally been coded in gesture research.

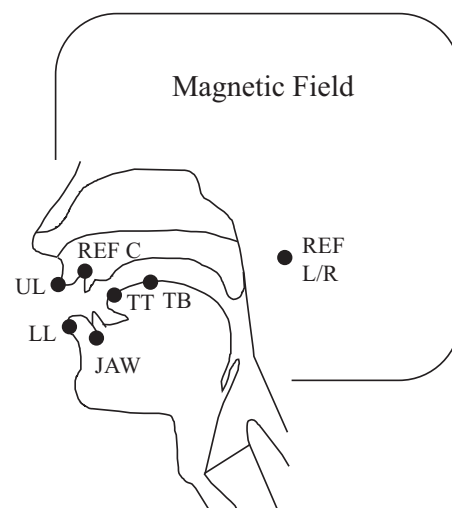


Fig. 1. EMA sensor placement schematic. EMA sensor coils were adhered at three points for head movement correction and five points to track articulator motion (lower incisor/jaw, tongue tip, tongue body, and upper and lower lips).

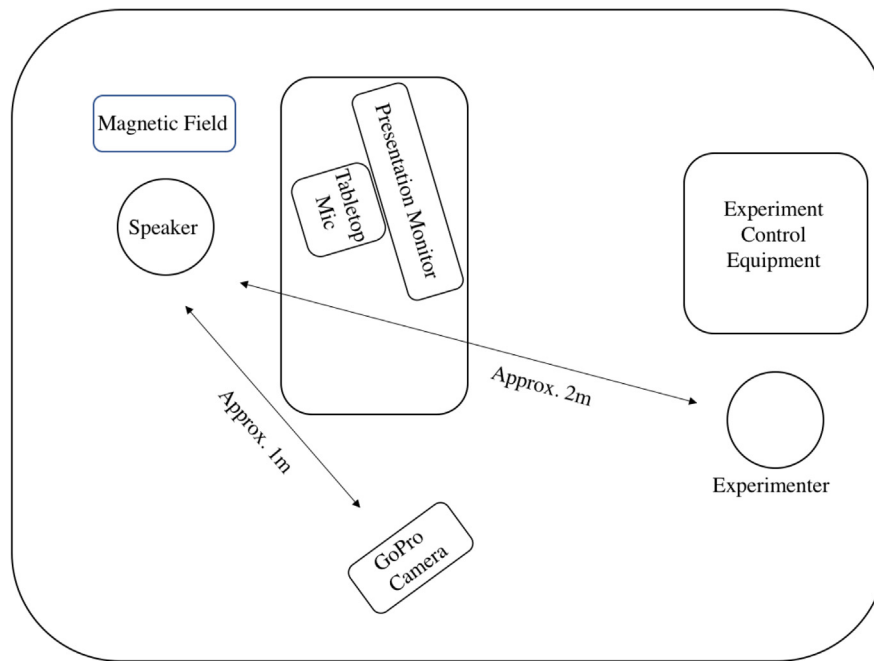


Fig. 2. Schematic of experimental setup with overhead view.

2. Methods

2.1. Subjects

Three native speakers of American English, all right hand dominant, participated in this experiment. Participants completed two experimental sessions, each lasting between 1.5 and 2 h, conducted between two days and ten days apart (only the first experiment session for each participant is analyzed here). The first of the three subjects had a slightly altered form of the preference task in the first experiment session, with open-ended questions rather than the more constrained preference questions described below. The re-formatted questions proved to elicit more speech content. All participants were naïve to the purpose of the study and were paid for their participation.

2.2. Experiment design & data acquisition

The experiment setup was consistent for the two experiment sessions completed by each participant. Participants were seated in an armless chair at a small table facing a computer monitor used to present instructions and stimuli. The participant was positioned next to a transmitter for the NDI (Northern Digital, Inc.) Wave² *electromagnetic articulograph* (EMA) system. EMA sensor coils were adhered to the right and left mastoid processes and the upper incisor to correct for head movement. To track articulator movements, sensor coils were adhered to the lower incisor, the tongue body (as centrally with respect to the oral length of the tongue as possible), the tongue tip (approximately 1 cm behind the anatomical tongue tip), and the upper and lower lips on the vermillion border. All sensors were adhered as close to the midsagittal plane as possible

(EMA sensor coil placement is shown in Fig. 1). Vocal tract articulator movements were sampled at 400 Hz.

A high-quality tabletop microphone was positioned near the participant, below the EMA magnetic field generator. Primary audio recordings were synchronized with the EMA system and recorded at 44.1 kHz. A GoPro video camera on a tripod was positioned approximately 1 meter diagonally to the right of and approximately 0.25 m above the seated speaker's head, recording video and secondary audio. This camera angle captured the participant's head and torso in the frame (see schematic in Fig. 2). Video was recorded at 29.97 frames per second with a resolution of 1920 × 1080 pixels; GoPro audio was simultaneously recorded in stereo at 48 kHz. The experimenter, acting as conversational confederate, was seated in the participant's line of sight at a distance of about 2 m, so the participant could direct responses to the experimenter. Audio, video, and movements of sensors affixed to positions in the vocal tract were recorded concurrently. After data collection was complete, the participant's EMA sensors were removed, the participant was paid, and the experiment concluded.

2.3. Stimuli

The experiment stimuli include twenty themes presented in two communication tasks, Demonstration and Preference (henceforth, *demo* and *pref* tasks, respectively). Stimuli were presented to the participant on the monitor using MARTA experiment control software (Tiede, Haskins Laboratories). The *demo* task was presented first, followed by the *pref* task; stimulus presentation order was randomized within task. Themes consist of common, gender-neutral activities that require the use of one's hands, such as buttoning one's shirt, opening an umbrella, or making a cup of tea. Five of the twenty

² <https://www.ndigital.com/msci/products/wave-speech-research/>

themes were analyzed for this experiment (see [Appendix A](#) for details).

The *demo* and *pref* tasks were both designed to elicit multi-modal speech with different qualities. We designed the *demo* task to elicit a high proportion of descriptive speech and manual gesture; traditional gesture researchers would likely call bodily movements in this task *representational/iconic* gestures ([Ferré, 2010](#); [Nobe, 2000](#)). The *pref* task was designed to elicit conversational speech and gesture that involves less visual description than the *demo* task; we expected that speakers would thus use movements referred to as *beat* gestures, which are not formally or semantically related to associated speech but which may be related to rhythmic or prosodic events in speech ([Krahmer & Swerts, 2007](#); [Leonard & Cummins, 2011](#)).

In the *demo* task, participants were asked to ‘Show me how you would x’, where x is the theme. The experimenter demonstrated a practice trial, on the theme ‘brushing your hair,’ describing the action step-by-step while simultaneously using manual gesture. Beyond the practice trial, participants were not provided explicit instructions regarding the use of manual gesture. Participants repeated the practice task to ensure that they understood the instructions. The *demo* task stimuli, consisting of the phrase ‘Please demonstrate how you x’ and accompanied by a black-and-white line drawing related to the theme, were presented to the participant on their desk monitor.



In the *pref* task, participants were asked to state their preference among two or more choices related to a theme and why they prefer their choice. For example, in the ‘brushing teeth’ theme, there are two questions: (1) Do you prefer an electric toothbrush or a regular toothbrush, and why? (2) Do you prefer whitening toothpaste or tartar control toothpaste, and why? Items were presented to the participant on the monitor with a line drawing related to the theme and a reminder to state and explain their preferences. [Table 1](#) describes the tasks and presents examples from the experiment.

2.4. Data processing

Raw EMA kinematic data was converted to TSV files using WaveFront software (Northern Digital, Inc.). This data was further processed for use in the custom EMA data analysis user interface Mview (Tiede, Haskins Laboratories), developed for MATLAB ([The MathWorks Inc., 2013](#)). The audio collected in sync with EMA data was annotated using TextGrids in Praat software ([Boersma & Weenink, 2016](#)), with information about speech epoch boundaries. Three categories of epoch, including speech, internal pause, and nonspeech, were annotated, but only *speech* epochs (periods of audible participant speech) were analyzed in this research.

GoPro video recordings were processed in QuickTime Player 7 video editing software ([Apple, 2010](#)). Trial length videos were resampled to 30 fps, 1280 × 720 px resolution in H.264 encoding (.mp4/.mov). Audio from each video file was converted to a mono-channel, 44.1 kHz .wav file. FlowAnalyzer software ([Barbosa & Vatikiotis-Bateson, 2013](#)) was used to create optical flow signals from video files. All movements were considered, including presumably non-communicative movements like touching the face or hair, to avoid making

Table 1
Experiment communication task descriptions.

Task type	Demonstration	Preference
Description	<ul style="list-style-type: none">• 20 themes• Participants describe their method for performing a basic task to the experimenter	<ul style="list-style-type: none">• 20 themes• 2–3 questions per theme• Participants respond to preference questions asked by experimenter
Example	Please demonstrate how you: Shuffle a deck of playing cards 	Please state which option you prefer and <i>why</i> you prefer that option over the alternative(s). Tea and coffee 

assumptions about gestural function. In each file, regions of interest (henceforth, ROIs) were selected for the right hand (RH), left hand (LH), and the head of the filmed participant. The selected ROIs include the entire range of movement for each body part in each trial ([Fig. 3](#) shows an example of ROI selection on a frame of recorded video). This method often caused the ROIs to overlap one another partially or completely, and the size of the selection regions thus differed by trial. For researchers wishing to apply these methods in their own research, we suggest a few design improvements: (1) use a ‘head-on’ camera angle in which each hand can be assigned to one half of the video frame (this speeds processing because ROIs can be reused across multiple trials); (2) have participants wear clothing that is easily distinguishable from their skin tone, and/or have participants wear brightly colored gloves that can be separated from the rest of the image in post-processing; and (3) use separate cameras or camera angles to record the movement of different ROIs.

Optical Flow signals created in FlowAnalyzer were processed in MATLAB. FlowAnalyzer creates three motion signals for each ROI: (1) x (horizontal): obtained by summing the x components of all velocity vectors inside the region; (2) y (vertical): obtained by summing the y components of all velocity vectors inside the region; and (3) mag (magnitude): obtained by summing the magnitude components (in polar coordinates) of velocity vectors inside the region. The magnitude signal gives the overall amount of motion within the region, ignoring direction.

The audio component from each segmented video trial was cross-correlated with the corresponding EMA audio using a Praat script ([Schlangen, 2014](#)) to align Optical Flow and EMA signals. MATLAB was used to create one-dimensional tangential velocity signals from Optical Flow and EMA motion data. In order to compare optical flow signals with other types of signals as we do here, it is necessary to resample any signals to be correlated to a common sampling rate (the ‘analysis’ rate). In this case, all tangential velocity signals were resampled to a common sampling rate of 200 Hz and were filtered using a ninth-order zero-phase Butterworth filter with a cutoff frequency of 3 Hz for Optical Flow signals ([Xiong & Quek, 2003](#)) and 12 Hz for EMA data. These two cutoff values were selected to smooth the different types of articulators, which

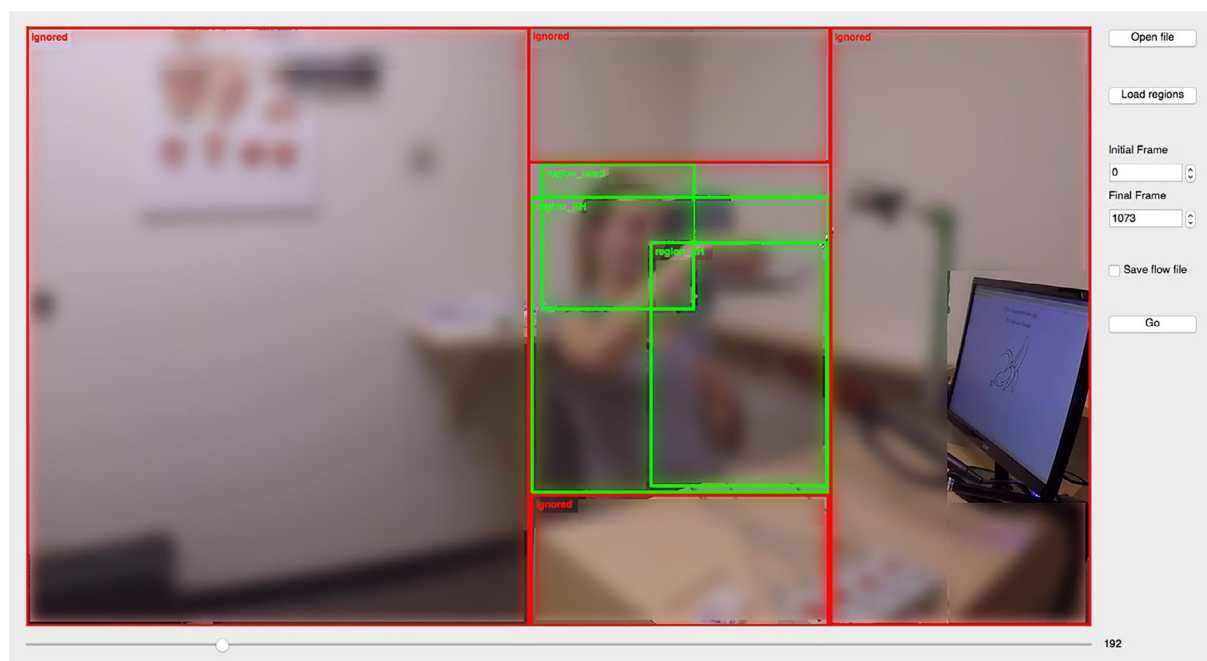


Fig. 3. Example ROI selection in FlowAnalyzer software. Green boxes correspond to individual regions and include the full range of movement for each body part in each trial. Areas of video frame outside of green boxes are ignored.

have distinct rates of characteristic oscillation. The data described and analyzed here is publicly available on Mendeley Data (Danner, Barbosa, & Goldstein, 2018).

2.5. Measures

2.5.1. Peak velocity

Peak velocity is used as an index of gestural magnitude for bodily movements and vocal tract articulators (Ostry et al., 1983). Tangential velocities are obtained by taking the first derivative of two- or three-dimensional position coordinates. These values are technically scalar values representing speeds, but we use the term ‘velocity’ here for ease of reference. The tangential velocities of the head, right hand, and left hand are described here. The times and magnitudes of tangential velocity maxima were obtained using MATLAB (The MathWorks Inc., 2013), which identifies a peak as a sample where velocity is at a local maximum compared to the samples on either side. Peak velocity is measured in millimeters per second for EMA signals, and pixels per frame for Optical Flow signals (1 frame is 0.005 seconds). Peak Velocity analyses compared kinematics during the *speech* epochs of the *demo* and *pref* conditions. We present descriptive statistics of peak velocities in the *demo* and *pref* tasks.

We also present the results of a trained classifier model for each participant, using right hand peak velocity magnitude as a predictor for determining whether the peak occurred during a *demo* or *pref* task. We divided each participant's data into a training set (80%) and a validation set (20%). We used a 10-fold cross-validation strategy for model parameter tuning, and we evaluated five common classifier algorithms for percentage accuracy and Cohen's kappa statistics. After selecting the preferred model for each participant, we evaluated the model's performance on the validation set for each participant.

2.5.2. Correlation Map Analysis

EMA and Optical Flow velocity time functions were analyzed with Correlation Map Analysis (CMA) software (Barbosa et al., 2012).³ CMA analyzes correlations between a pair of time-varying signals at a range of temporal offsets in either direction (i.e., Signal 1 can be either leading or lagging Signal 2). CMA is a signal analysis technique, previously used in articulatory phonetics and multimodal speech research, to assess the degree of similarity or coordination between two signals. Similar signal comparison tools include cross-correlation analysis (Paxton & Dale, 2013), cross-recurrence analysis (Louwerse, Dale, Bard, & Jeuniaux, 2012), wavelet analysis (Xiong & Quek, 2006), functional data analysis (Parrell, Lee, & Byrd, 2013), etc. An extensive review of how CMA compares to other methods can be found in the article introducing CMA (Barbosa et al., 2012). We use CMA here because it offers several advantages in multimodal speech analysis applications, like the ability to precisely quantify correlation instantaneously *and* at a range of lags, which is crucial due to the lack of strict temporal coordination among biological signals (Winfree, 2001). Another advantage of CMA is the mapping component, which visualizes the pattern of changes in correlation between two signals over a user-selected time period, including information about whether correlation is positive or negative. Correlation values produced by CMA are dependent on three user-defined parameters: (1) the forgetting factor, η , used to compute the correlation at any point in time, which affects the granularity of the correlation measure (effectively, the window size over which correlation is computed); (2) the temporal range of between-signal delays over which correlation is calculated; and (3) the use of a uni-directional or a bi-directional filter, which specifies whether only

³ The MATLAB implementation of CMA software is freely available at <https://github.com/avspeech/cma-matlab>

past samples (uni-directional) or both past and upcoming samples (bi-directional) are used to compute correlation.

In the present experiment, CMA analysis was performed across an entire trial, irrespective of speech epoch, comparing the EMA-derived *jaw* velocity signal and the Optical Flow-derived *right hand* (RH) velocity signal. A ± 1 s delay range is used to compute correlations at every offset of the signals within that range. Signals were analyzed bi-directionally, because speech kinematic movements depend on information both preceding and following the target movement (Barbosa et al., 2012); and the forgetting factor was set to a small value, $\eta = 0.030$, in order to support a fine-grained analysis on a time-scale appropriate for phonetic research.

In the results section below, we qualitatively discuss the occurrence of two individual manual movements and their respective relationships to accompanying speech, one of which occurs in the *demo* task, and the other of which occurs in the *pref* task. The correlation matrix produced by CMA is also used to produce a quantitative measure, which is the probability that the two signals' correlation is above a threshold value ($\rho = 0.5$), normalized by the length of a trial (Fuhman, 2014). We refer to this measure as *PrPosCorr*, the Probability of above-threshold Positive Correlation between the two signals of interest. We looked at correlations between these two signals at a range of temporal offsets between signals: when the right hand signal is delayed with respect to the jaw signal by 50 ms, 100 ms, 200 ms, 500 ms and 1000 ms, when the jaw signal is delayed with respect to the right hand signal by 50 ms, 100 ms, 200 ms, 500 ms, 1000 ms, and at zero delay. *PrPosCorr* is used to study the questions of whether likelihood of cross-modal correlation serves to distinguish the two communicative tasks employed here and whether correlation varies dependent on temporal offset values.

3. Results

The results described below comprise data from five out of twenty elicited themes (laundry, jars and containers, umbrellas, bananas, and candy) for which a full data set was available for all speakers (some data loss occurred due to video and/or EMA recording errors). Descriptive analysis, visualization, and classifier model training is performed in R version 3.3.3. (R Core Team, 2017) using packages *psych* v1.7.3 (Revelle,

2017), *ggplot2* v2.2.1 (Wickham, 2009), and *caret* v6.0-78 (Kuhn et al., 2017).

3.1. Peak velocity

3.1.1. Hands & head peak velocity distributions by task

The *n* value, mean, min, max, and peaks per second (PPS) shown in Table 2 for each of the three body movement regions were calculated over all velocity peaks occurring during speech epochs. Because trials differed in length within and between tasks and speakers (and *demo* trials tended to be shorter than *pref* trials), we used mean peaks per second as a measure to evaluate the average frequency of velocity peaks for each articulator among the different speakers and tasks. These descriptive statistics demonstrate that for all speakers and movement regions, average peak velocity values were higher in the *demo* task than in the *pref* task. The mean PPS values for the head and hands were also higher in the *demo* task than in the *pref* task. For speakers Pilot and M1, the average head peak velocity magnitudes for all participants were lower, in both tasks, than were the average magnitudes for the hands, while for F1, the average head peak velocity magnitude was lower than that of the right hand but greater than that of the left hand. Somewhat unexpectedly, the mean of left hand velocity magnitudes for participants Pilot and M1 was greater than the mean of right hand velocity magnitudes; this outcome may have been influenced by the overlapping ROIs described in Section 2.4. The histograms in Figs. 4, 5 and 6 show more detail regarding distribution of peak velocity magnitudes by task; for example, there are typically more very low-magnitude tokens in the *pref* task than in the *demo* task for the right and left hand, whereas peak velocity magnitudes in the *demo* task are more dispersed across a range of bins. The head appears to have more overlap of magnitude distributions across the two tasks, particularly for the participants Pilot and M1.

3.1.2. Models for classifying right hand velocity peaks by communicative task

We analyze velocity peaks of the right hand occurring during the *speech* epoch separately for each speaker. We tested several common classifier algorithms, including Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), K-nearest neighbors (KNN), Support Vector Machines

Table 2

Peak velocity magnitude (in pixels/frame) for head and hands during speech epoch by communicative task and participant. Mean peak velocity magnitudes are greater in *demo* than *pref* tasks for all speakers and all movement regions. Mean peaks per second (PPS) was higher in the *demo* task than the *pref* task for all speakers and all movement regions.

		<i>n</i>	Demo task			PPS	<i>n</i>	Pref task			PPS
			Mean (SD)	Min	Max			Mean (SD)	Min	Max	
Right hand	Pilot	214	0.78 (0.55)	0.11	2.83	1.59	219	0.63 (0.52)	0.11	2.64	0.93
	M1	165	0.86 (0.54)	0.31	3.09	1.31	192	0.48 (0.13)	0.26	0.99	0.81
	F1	215	1.07 (0.79)	0.22	4.1	1.17	302	0.4 (0.54)	0.06	4.51	0.80
Left hand	Pilot	210	0.88 (0.71)	0.09	4.35	1.58	226	0.66 (0.65)	0.1	3.54	0.96
	M1	160	0.94 (0.58)	0.31	3.28	1.26	209	0.51 (0.25)	0.3	2.12	0.88
	F1	215	0.94 (0.74)	0.16	4.52	1.16	318	0.28 (0.42)	0.04	4.01	0.84
Head	Pilot	200	0.5 (0.41)	0.15	2.74	1.50	233	0.46 (0.31)	0.14	2.58	0.99
	M1	166	0.53 (0.19)	0.31	1.88	1.30	213	0.46 (0.11)	0.23	1.31	0.90
	F1	212	0.72 (0.78)	0.24	8.12	1.13	299	0.39 (0.22)	0.07	1.7	0.79

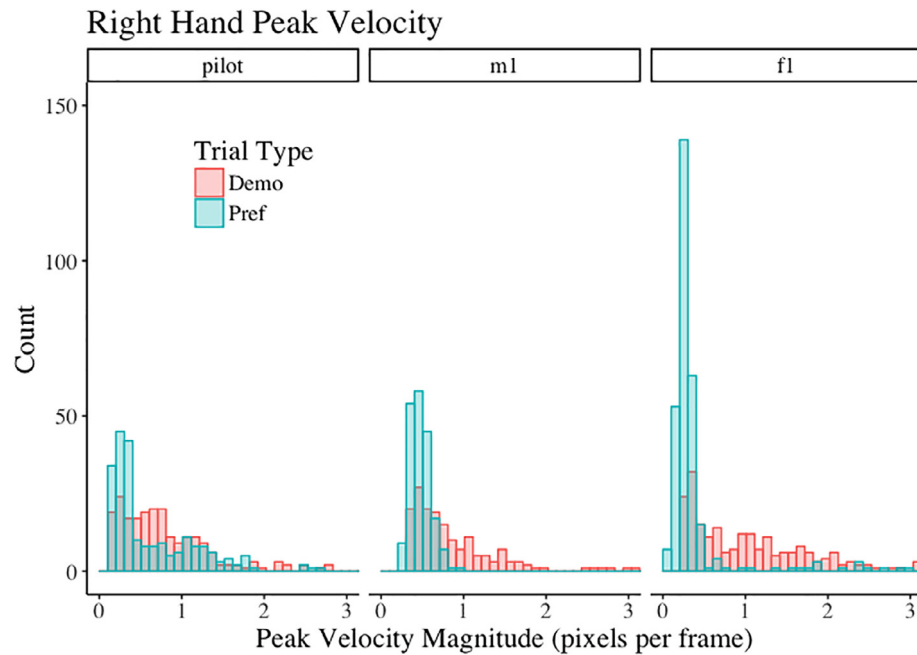


Fig. 4. Histogram of right hand peak velocity by task and participant (bin width determined by Freedman-Diaconis rule). Speakers' right hand velocity peak magnitudes in the *pref* task contain several low-magnitude tokens, whereas velocity peak magnitudes in the *demo* task are more dispersed across a range of bin sizes.

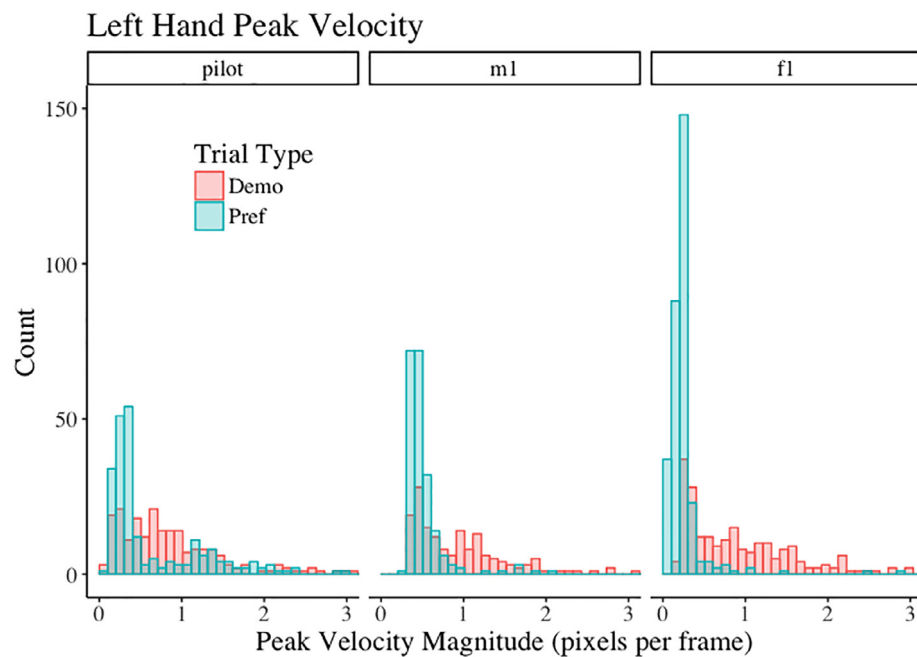


Fig. 5. Histogram of left hand peak velocity by task and participant (bin width determined by Freedman-Diaconis rule). Speakers' left hand velocity peak magnitudes in the *pref* task contain several low-magnitude tokens, whereas velocity peak magnitudes in the *demo* task are more dispersed across a range of bin sizes.

(SVM) with Radial Basis Function (RBF) Kernel, and Random Forest (RF).⁴ The two statistics we report in Table 3, raw % Accuracy and Cohen's kappa (a statistic describing the ratio of observed versus expected accuracy), are averaged over the ten resampled test datasets used for each algorithm. The classification algorithm with the highest % accuracy and Cohen's

kappa statistics for all participants, denoted in Table 3 with bold text, was the SVM model. We therefore used SVM to classify the validation datasets, which gives the classification accuracy statistics achieved for each speaker.

The results in Table 4 show that classification accuracy, even given only one predictor (peak velocity magnitude), is decent for predicting the communicative context during which that peak occurred. The raw % accuracy of the classifier was fairly high (~80%) for participants F1 and M1, but lower for

⁴ The R code used for these analyses is available in our published dataset (Danner et al., 2018), in 'Jphon_revised_analysis_v4.r' in the Analysis Scripts folder.

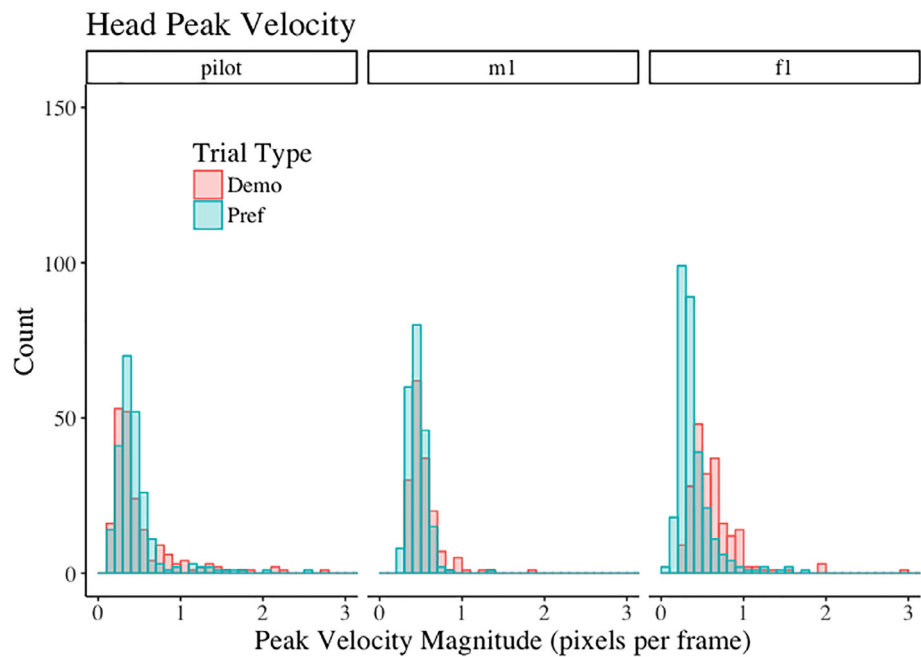


Fig. 6. Histogram of head peak velocity by task and participant (bin width determined by Freedman-Diaconis rule). Speakers’ head velocity peak magnitudes in the *pref* task contain several low-magnitude tokens, whereas velocity peak magnitudes in the *demo* task are more dispersed across a range of bin sizes.

Table 3
Mean % accuracy and mean Cohen’s kappa (averaged from 10-fold cross-validation) for five classification algorithms applied to determine task classification for right hand velocity peaks, by participant. The bold values represent the highest % Accuracy and Cohen’s kappa values attained for each speaker. The SVM classifier had the highest % accuracy and Cohen’s kappa values for all speakers.

		% Accuracy	Cohen’s kappa
LDA	Pilot	0.55	0.10
	M1	0.73	0.43
	F1	0.75	0.44
CART	Pilot	0.62	0.24
	M1	0.71	0.40
	F1	0.78	0.53
KNN	Pilot	0.58	0.15
	M1	0.71	0.40
	F1	0.8	0.58
SVM	Pilot	0.65	0.29
	M1	0.74	0.45
	F1	0.81	0.61
RF	Pilot	0.58	0.15
	M1	0.68	0.36
	F1	0.74	0.47

the pilot participant (however, the ~60% accuracy for the Pilot was still slightly better than the accuracy of the model with no information, at ~51%). With more than one predictor, it is reasonable to expect that classification accuracy would improve beyond what is possible with one predictor.

Table 4
SVM model accuracy descriptions by participant (% accuracy is raw classification accuracy, Cohen’s kappa is the ratio of observed versus expected accuracy, and the No Information Rate is a measure of accuracy when no additional information is given beyond the distribution of data in each class. % accuracy was greater than the No Information Rate for all speakers. For speakers F1 and M1, % accuracy reached approximately 80% with just one predictor.

		% Accuracy (95% CI)	Cohen’s kappa	No information rate
SVM classifier	Pilot	0.6 (0.49, 0.71)	0.20	0.51
	M1	0.79 (0.68, 0.88)	0.56	0.54
	F1	0.80 (0.71, 0.87)	0.57	0.58

3.2. Correlation Map Analysis

3.2.1. Qualitative analysis

To highlight the capabilities of CMA applied to multimodal speech research, we describe two examples of manual gestures in multimodal speech using correlation maps from our experimental data. In addition to the correlation map, these examples show a frame from the video recording of the trial, and the same frame from an optical flow movie, to provide a qualitative analysis of an exemplary gesture in each of the *demo* and *pref* conditions (optical flow movies can be created using the `create_flow_movie.py` Python script distributed with the FlowAnalyzer software).

The example in Fig. 7 is drawn from the Pilot participant’s data in the *demo* task; the theme for this example is “bananas” (see Appendix A). The video frames show the manual gesture of interest, which is also visible as a peak (the origin of the two red arrows) in the “Right hand Velocity” panel. The optical flow movie represents regions where Flow Analyzer has detected movement. The magnitude of the movement is encoded as darkness in the optical flow movie – the greater the magnitude of pixel displacement, the darker that region of the video frame will appear. This gesture is produced as the speaker utters the phrase “Once I get to the very end. . .” The speaker is referring to peeling a banana, and this gesture, which fits the description of a ‘representational’ gesture because of its visually expressive

hand shape, appears to mime a banana held in the participant's right hand.

The lower set of panels in Fig. 7 shows the data displayed in the Correlation Map graphical user interface (the x-axis, representing time, is shared across all panels in the correlation map GUI). The panel labeled “audio waveform” is the acoustic waveform of the speech during this sample. The next panel down shows Signal 1, which in this case represents the right hand Optical Flow motion signal during the sample. The next panel, labeled “Jaw Velocity,” shows Signal 2, the movement velocity acquired from the EMA sensor adhered to the speaker's Jaw. In the “Instantaneous Correlation” panel, a black line shows the continuous correlation between our two signals of interest, which we use as a measure of the coordination between speech (represented by jaw motion) and gesture (represented by right hand motion). The center of the Instantaneous Correlation panel represents zero correlation, while the top represents positive correlation and the bottom represents negative correlation. The bottom panel, labeled “2D correlation map” shows all the CMA-computed correlations between the right hand and jaw: the vertical center of the map represents the instantaneous correlation between the two signals. The top half of the map represents the part of the temporal delay range in which the jaw signal is delayed with respect to the right hand, and the bottom half of the map represents the part of the temporal delay range in which the right hand signal is delayed with respect to the jaw signal (the temporal delay in this example is ± 1 s). The darkest red colors indicate strong positive correlation, while the dark blue colors indicate strong negative correlation; green indicates no correlation.

In Fig. 7, the correlation between right hand and jaw movement in the *demo* task varies over time and is neither

consistently high nor consistently low in this sample; a region of strong positive correlation occurs between 29 and 30 s (visually represented by a red region), and later a blue region of strong negative correlation occurs between 32 and 34 s. Considering the vertical extremes of the map, both strong positive and negative correlations are observed when the signals are compared at lags (e.g., the dark blue region of negative correlation shown at the bottom of the map between 29 and 30 s, when the right hand signal is delayed with respect to the jaw, or the red region of positive correlation at the top of the map, around 29 s, in which the jaw signal is delayed with respect to the right hand). We can also observe that the jaw signal contains many more peaks than the right hand signal in this sample, indicating that jaw position changes more rapidly than hand position.

Fig. 8 shows a multimodal speech snippet with an example of a manual gesture produced by participant F1 in the *pref* condition. The theme in this example is again “bananas” (see Appendix A), and the example gesture is produced during the utterance “...part of a fruit bowl...” This example of manual gesture is quite different from the gesture in Fig. 7. For example, the speaker's hand is mostly out of sight (all participants rested their hands in their lap for the majority of *pref* task trials), with only the thumbs of each hand emerging into frame. The hands do not appear to depict anything with form/shape or location, which indicates that this gesture might best be described as a ‘beat’ gesture in the McNeill typology (McNeill, 1992). The right hand velocity signal does not show as many steep velocity peaks as the same signal in Fig. 7. The instantaneous correlation panel shows primarily near-zero and small negative correlations between right hand and jaw velocity, but the relationship between the signals fluctuates over time, as observed in the previous example. The correlation

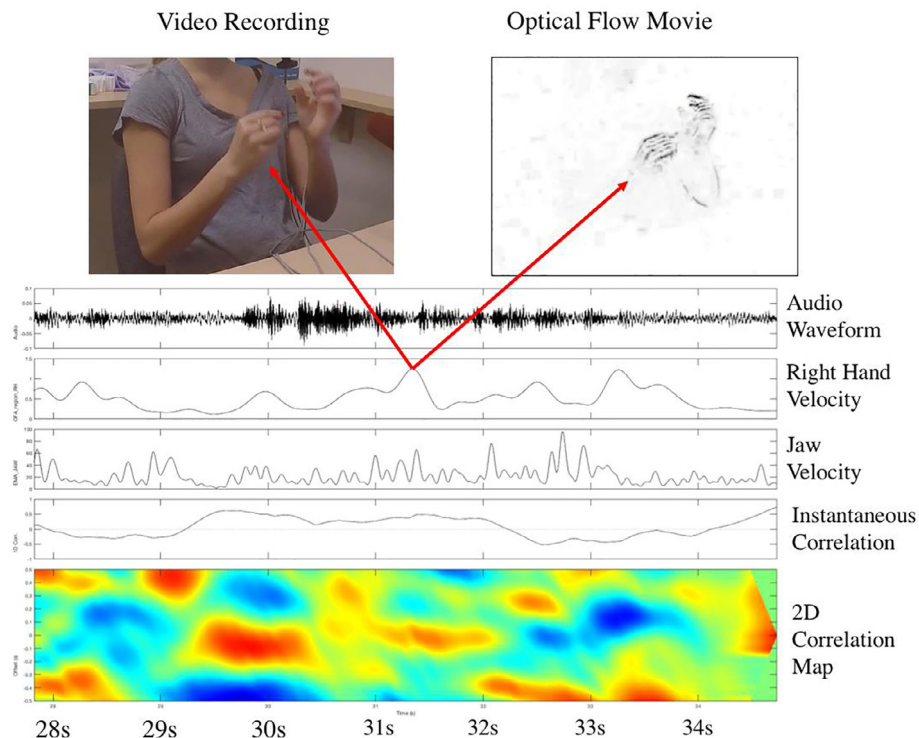


Fig. 7. Example *demo* task gesture shown in a video frame, in an optical flow movie frame, and in a correlation map. The right hand velocity panel shows fewer peaks in comparison to the jaw velocity panel. The bottom panel demonstrates that the degree and polarity of correlation between the right hand and jaw velocity signals fluctuates over time and at various delays between the signals.

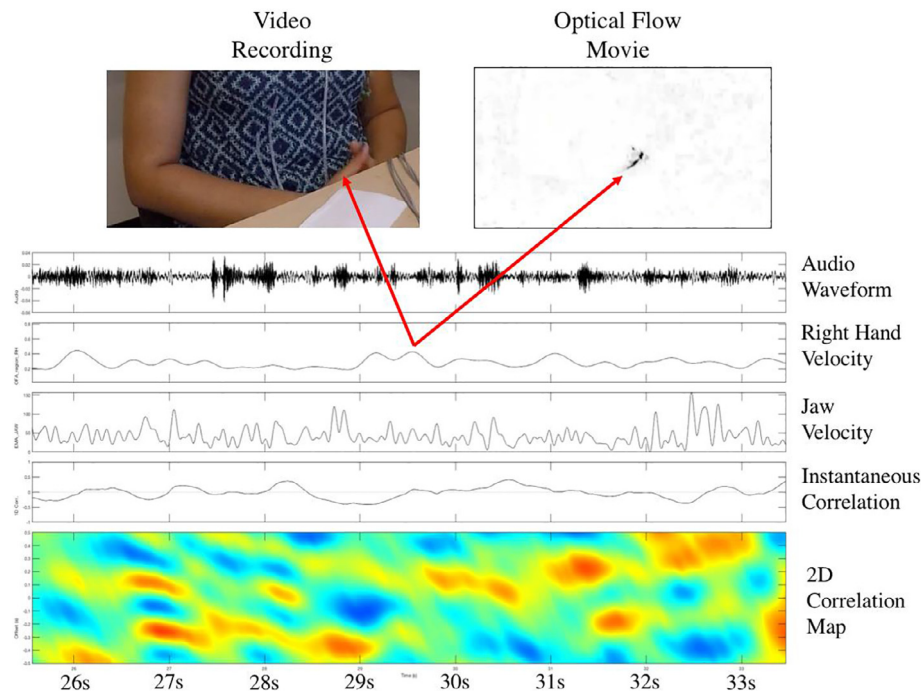


Fig. 8. Example *pref* task gesture shown in a video frame, in an optical flow movie frame, and in a correlation map. The velocity peaks in the right hand signal are generally smaller than in the *demo* example (Fig. 4), and the instantaneous correlation and correlation map show fewer regions of strong positive or negative correlation between right hand and jaw velocity than in the *demo* example.

Table 5
PrPosCorr at a range of temporal offsets, by task and participant. PrPosCorr mean values are generally greater in *pref* tasks than *demo* tasks. Greater mean PrPosCorr values in *pref* versus *demo* tasks only occur at short temporal offsets between right hand and jaw velocity, demonstrating a possible *gesture lead* effect.

		Demo task			Pref task		
		Mean (SD)	Min	Max	Mean (SD)	Min	Max
−1000 ms	Pilot	0.04 (0.06)	0.01	0.14	0.04 (0.02)	0.01	0.08
	M1	0.05 (0.01)	0.04	0.06	0.04 (0.02)	0.01	0.06
	F1	0.02 (0.02)	0	0.04	0.02 (0.02)	0	0.05
−500 ms	Pilot	0.06 (0.03)	0.03	0.11	0.05 (0.01)	0.03	0.07
	M1	0.02 (0.02)	0	0.04	0.04 (0.02)	0.02	0.07
	F1	0.02 (0.01)	0	0.03	0.03 (0.01)	0.02	0.04
−200 ms	Pilot	0.04 (0.03)	0	0.07	0.04 (0.01)	0.03	0.06
	M1	0.05 (0.04)	0.01	0.1	0.03 (0.02)	0	0.06
	F1	0.04 (0.03)	0	0.09	0.04 (0.02)	0.03	0.07
−100 ms	Pilot	0.07 (0.04)	0.02	0.12	0.07 (0.01)	0.05	0.09
	M1	0.03 (0.01)	0.02	0.05	0.05 (0.02)	0.02	0.09
	F1	0.02 (0.01)	0	0.03	0.06 (0.02)	0.04	0.1
−50 ms	Pilot	0.06 (0.04)	0.02	0.11	0.09 (0.02)	0.05	0.1
	M1	0.05 (0.02)	0.03	0.08	0.08 (0.03)	0.04	0.12
	F1	0.02 (0.02)	0	0.04	0.06 (0.03)	0.04	0.09
Zero offset	Pilot	0.05 (0.04)	0.01	0.1	0.1 (0.03)	0.06	0.15
	M1	0.05 (0.02)	0.03	0.08	0.09 (0.03)	0.05	0.13
	F1	0.04 (0.03)	0.01	0.08	0.08 (0.05)	0.03	0.14
+50 ms	Pilot	0.04 (0.03)	0.02	0.1	0.1 (0.05)	0.05	0.17
	M1	0.03 (0.02)	0.01	0.05	0.11 (0.04)	0.04	0.14
	F1	0.04 (0.03)	0.01	0.07	0.07 (0.05)	0.03	0.14
+100 ms	Pilot	0.04 (0.03)	0.01	0.09	0.11 (0.05)	0.06	0.18
	M1	0.03 (0.02)	0	0.06	0.12 (0.04)	0.05	0.16
	F1	0.05 (0.04)	0.01	0.09	0.07 (0.04)	0.02	0.12
+200 ms	Pilot	0.03 (0.03)	0	0.06	0.07 (0.04)	0.05	0.14
	M1	0.05 (0.03)	0	0.09	0.1 (0.05)	0.04	0.18
	F1	0.03 (0.02)	0.01	0.06	0.05 (0.02)	0.03	0.08
+500 ms	Pilot	0.04 (0.03)	0.01	0.1	0.05 (0.02)	0.02	0.08
	M1	0.03 (0.02)	0.01	0.06	0.05 (0.02)	0.02	0.08
	F1	0.02 (0.01)	0	0.04	0.03 (0.01)	0.01	0.05
+1000 ms	Pilot	0.04 (0.02)	0.01	0.06	0.05 (0.02)	0.01	0.07
	M1	0.04 (0.04)	0.01	0.1	0.04 (0.02)	0.01	0.06
	F1	0.02 (0.01)	0.01	0.03	0.03 (0.01)	0.02	0.04

map shows a few small ‘bursts’ of strong correlation across lags, but fewer sustained regions of strong correlation than in the Fig. 7 example. A striped region around the 27 s mark shows alternating strong positive and strong negative correlation between modalities, dependent on the offset value. The example discussed in Fig. 8 has several qualities that distinguish it from the multimodal speech produced during the *demo* task; namely, qualitative differences in the manual gesture, as well as differences in the speech-gesture relationship, as shown by distinct patterns in the correlation map.

3.2.2. PrPosCorr

In the PrPosCorr measure, negative offset values (e.g., –1000 ms, –50 ms) indicate a delay of the first sample of signal 1 (right hand velocity) with respect to the first sample of signal 2 (jaw velocity), and positive offset values (e.g. +50 ms, +1000 ms) indicate a delay of signal 2 (jaw velocity) with respect to signal 1 (right hand velocity). The PrPosCorr measure is effectively a percentage likelihood of strong positive correlation for each offset value across an entire trial, normalized by the length of the trial. Each mean is taken from five analyzed trials performed in each task for each participant.

The mean PrPosCorr values in Table 5 demonstrate that, in most cases, each participant had a higher mean PrPosCorr value during the *pref* task than during the *demo* task (exceptions to this observation occurred for the Pilot participant at –1000 ms, –500 ms, –200 ms, and –100 ms; for participant M1 at –1000 ms, –200 ms, and +1000 ms; and for participant F1 at –200 ms and –1000 ms). These results imply that strong positive correlation between speech and manual gesture – as represented by jaw velocity and right hand velocity, respectively – is more common when speakers participate in a conversational communicative task (like the *pref* task) than when speakers are performing a descriptive task (like in the *demo* task). When considering the delay values, it is evident that for most participants, delays between ± 200 ms in the *pref* task show the highest probability of positive correlation. In particular, high PrPosCorr means are observed in the *pref* task at zero offset, and at +50 ms and +100 ms offsets (representing offsets at which the jaw signal is delayed with respect to the right hand signal). This supports the previous observations of a ‘gesture lead’ effect (Wagner et al., 2014), in which correlation between speech and gesture signals is highest at a short (~ 100 ms) delay of speech with respect to manual gesture. At longer delays, particularly ± 200 ms or more, PrPosCorr averages level off and differences in the speech-gesture relationship by task are less apparent. This quantitative measure of the coordination between speech and gesture supports the task-based distinction noted in the qualitative assessment conducted in Section 3.2.1, but additional detail is provided by the precision of the PrPosCorr measure at different offsets between signals.

4. Discussion

4.1. Using quantitative and coordinative measures to characterize manual gesture and the Speech-Gesture relationship

In this paper, our goal was to investigate whether quantitative characteristics might be used to distinguish multimodal

speech behaviors dependent on communicative task/context. We proposed that if kinematic and coordinative properties of multimodal speech are distinctive, then it may be possible to consider a systematic relationship between quantitative properties of manual gesture and the qualitative annotations of manual gesture prevalent in gesture and multimodality research. Establishing this relationship would allow researchers to standardize the detection and description of manual gestures, which in turn could decrease the time cost inherent to multimodal speech research and improve researchers’ ability to generalize across research results in this field. Although typical spontaneous speech rarely conforms to the strict experimental manipulations of communicative context we have used in this research, our findings suggest that some communicative contexts show distinctive kinematic and coordinative properties. By implementing the gesture detection and multimodal speech quantification techniques that we have described here, our aim is to give linguistic researchers an additional tool that will make the corroboration of different lines of gesture research achievable, and which could open new lines of questioning in linguistically-informed multimodal speech research.

To study whether communicative tasks can be distinguished quantitatively, we devised the Demonstration (*demo*) task, to elicit descriptive language and a high proportion of descriptive manual gestures that bear a formal or semantic relationship to accompanying speech (what some researchers have termed ‘representational’ gestures); and the Preference (*pref*) task, a conversation-oriented task, in which we expected speakers to produce a higher proportion of non-descriptive gestures without a specific formal or semantic relationship to the accompanying speech (what some researchers have termed ‘beat’ gestures). We demonstrated that peak velocity magnitude, as the sole predictor in a classification task, classified manual gesture according to communicative task with between 60% and 80% accuracy, depending on the participant. Our other quantitative findings show that hand and head peak velocity magnitudes tend to be higher in the *demo* task compared to the *pref* task. Histograms of peak velocity magnitudes by speaker and movement region also demonstrated somewhat different distributions by communicative task. For both hand and head movement regions, speakers also exhibited distinctions in the number of velocity peaks per second by communicative task. Whereas manual gesture typologies in previous research have focused on observation of formal/semantic properties of individual gestures to determine the communicative intent of a given gesture (Hostetter et al., 2007; McNeill, 1992), here we show that quantitative properties are also capable of distinguishing multimodal speech in communicative tasks in a way that could be used to assess the relationship between properties of bodily movement and quantifiable aspects of linguistic processes like speech planning, prosody and semantics.

We also use CMA to assess the time-varying coordination between speech and manual gesture in our two communicative tasks. The variety of evidence pertaining to the relationship between speech and gesture has long vexed the multimodal speech research community because different annotation schemes and conflicting findings on the timing of speech and gesture make it difficult to explicitly connect speech and gesture as two pieces within a larger communicative system

(de Ruiter, 2000; McNeill, 1992; Rusiewicz, 2011). In addition to the qualitative differences observed in patterns of correlation within the correlation maps described for each task, we demonstrated, using the *PrPosCorr* measure, that speech and manual gesture show stronger positive correlations in the *pref* task than in the *demo* task, and strong positive correlations are especially likely to occur when speech and gesture signals are time-aligned or slightly offset (± 200 ms or less). This offset value is consistent with previous observations about the delay between speech and manual gesture movements (Krivokapić et al., 2017). We take our *PrPosCorr* finding as early quantitative evidence that task- and offset-based distinctions in cross-modal coordination are communicatively meaningful, and that speakers may have cognitive control over cross-modal coordination beyond the fortuitous entrainment that regularly occurs in many physical and biological systems (Turvey, 1990; Winfree, 2001).

One potential influence on gesture quantification not explicitly considered in our study is the presence of an audience. In the *demo* task, participants were asked to demonstrate an activity to the experimenter, and in the *pref* task, participants engaged in conversational turn-taking with the experimenter, but crucially, both tasks were performed in the experimenter's presence. We expected that the physical presence of the experimenter would encourage participants to use gestures, and use them in a naturalistic way, based on evidence from earlier research that face-to-face interaction acts as a 'default' communicative setting (Bavelas, Gerwing, Sutton, & Prevost, 2008; Garrod & Pickering, 2004). In an experiment investigating the role of audience presence and visibility on the rate and size of manual gesture, Bavelas et al. (2008) showed that gesturing rate and magnitude in a picture description task was greatest in a face-to-face dialogue, lower in an over-the-phone dialogue, and lower still when participants gave a 'monologue' to a tape recorder. Several previous quantitative studies of gesture have used controlled experimental settings to elicit gesture with limited face-to-face interaction with an interlocutor (Krahmer & Swerts, 2007; Krivokapić et al., 2017; Rochet-Capellan et al., 2008), and given the findings of Bavelas et al., we suspect it would be difficult to directly compare our observations of gesturing rate and magnitude to those of other studies because of this distinction in audience presence/interaction. It may be interesting to consider in future research whether the task-based distinctions we found in quantitative measures of gesture hold up when other factors like audience presence/absence and visibility are systematically controlled.

Consideration should also be paid to the topic of head gesture and other non-manual bodily gesture. Although we didn't experimentally manipulate co-speech head movement, we observed that properties of head movements, like hand movements, might be dependent on communicative context. Research comparing the effects of different kinds of non-speech gestures (hand gestures, head nods, and eyebrow movements) concluded that each kind of non-speech gesture had a similar effect on the production and perception of speech prominence (Krahmer & Swerts, 2007), and the research supports a basic equivalence among different kinds of non-speech gestures. This equivalence is likely advantageous given the

variety of contexts in which humans communicate (Bavelas et al., 2008): we can speak fluently even when our hands are occupied in holding or carrying objects, and we might use non-manual bodily gestures or facial expressions to aid communication when we are trying to communicate in difficult environments. It is our hope that by applying the analytic techniques we have presented here, this claim of equivalence among different non-speech gestures can be supported with more quantitative, empirical data.

4.2. Review of methods for multimodal speech data collection and analysis

We have obtained promising results using a novel combination of data collection and quantification techniques for multimodal speech, but improvements and modifications to these techniques for different experimental applications need to be considered. We have demonstrated that optical flow can be used to quantify bodily movement on its own or in conjunction with other motion capture equipment. In our methods section, we justified our choice to use overlapping ROIs for the hands and head because of the desire to capture as much movement as possible, yet this decision made analysis of movement in different regions more difficult to attribute to a given body part. Depending on the applications or research goals, ROI selection, as well as the positioning of recording equipment, should be carefully considered when designing multimodal speech experiments (Barbosa et al., 2008). Techniques like color separation or pose estimation should also be considered in future research (Danner, 2017; Simon et al., 2017). These alternative optical flow analysis techniques can ultimately be compared to measures from data gloves or similar methods for obtaining kinematic data to gain a better understanding of co-speech hand and arm kinematics (Wagner et al., 2014).

The quantitative measures used in this experiment, such as peak velocity magnitude, PPS, and *PrPosCorr*, produced encouraging results using tangential velocities derived from EMA and Optical Flow motion signals. These measures demonstrate that information about communicative environment can be obtained from quantitative properties of manual gestures and the coordinative properties of speech and gesture, which may ultimately enrich existing observational descriptions of the form and proposed function of bodily gesture in multimodal speech. Still, these and other quantitative measures such as directionality of movement and time to peak velocity or inter-peak intervals may also prove to be useful measures in future multimodal speech and gesture research (Byrd, 1996; Krivokapić et al., 2017).

We used CMA to assess time-varying coordination in this experiment, in contrast to previous approaches using, e.g., anchor points to assess instantaneous temporal coordination across modalities. Our CMA method showed evidence of a systematic relationship between communicative task and likelihood of correlation between modalities, and it also demonstrated that assessing imprecise temporal coordination provides valuable insights regarding multimodal speech. Future CMA investigations might also consider breaking trials into smaller pieces such as speech epochs or regions around

prosodic boundaries, to understand whether correlation across modalities systematically varies at points or regions within an utterance (Danner, 2017; Wagner et al., 2014).

Ultimately, the techniques and analyses presented here are aimed at expanding the data collection and analysis options available to gesture researchers and multimodal speech researchers in linguistics. We have shown that our techniques make it possible to semi-automatically detect the occurrence of a gesture by locating velocity peaks. Further, our techniques allow for communicative contexts to be distinguished on the basis of kinematic and coordinative properties of multimodal speech. These techniques highlight the possibilities of using quantitative characteristics of multimodal speech to validate and standardize existing descriptions of multimodal speech, and to do so in spontaneous/naturalistic communicative contexts. With additional research, it may even be possible to understand ‘phonetic’ details of speech-accompanying gesture, like temporal and kinematic properties, and semantic/communicative functions of speech-accompanying gesture in a systematic way, a possibility many researchers have suggested (Abner, Cooperrider, & Goldin-Meadow, 2015; Wagner et al., 2014). With the techniques discussed in this paper, we aim to present a new way of thinking about the characterization of manual gesture in terms of communicative context, which will enhance future research in multimodal speech.

5. Conclusions


The techniques for multimodal speech data collection and analysis presented here offer a method for improving and expanding the linguistic study of multimodal speech. While considerable advances have been made in the study of multimodality over several decades of research (Wagner et al., 2014), the commonly used technique of coding gestures manually limits the quantity of data that can be analyzed, leads to uncertainty in consistency of coding across studies, and limits the kind of quantitative analyses that can be performed. This paper describes an affordable and accessible kinematic data collection method called FlowAnalyzer (Barbosa & Vatikiotis-

Bateson, 2013), that does not require special motion capture equipment, and which can be applied to existing digital video. Though the method may not be appropriate for all research applications, it improves upon previous video-based multimodal speech research because it offers the ability to quantify movement magnitude and direction. This paper also presents a technique for detecting manual gestures in optical flow signals using velocity peaks, which can be used to validate human-annotated multimodal speech data. Future research should conduct a thorough comparison between our automatic gesture detection methods and common human annotation methods to determine how comparable these techniques are. Finally, this paper describes how quantitative measures, including time-varying correlation measures obtained from CMA (Barbosa et al., 2012), may be appropriate for investigating some kinds of multimodal speech research questions. The results of our experiment suggest that communicative tasks can be distinguished based on the kinematic and coordinative characteristics of multimodal speech within those tasks. More research is needed to determine the persistence of these effects across a wide variety of communicative contexts, but these initial results indicate that quantitative and time-varying analysis of multimodal speech can provide valuable research insights. It is possible that further use of these methods can lead to a more general way of identifying and describing speech-accompanying gestures in multimodal speech research, using both qualitative and quantitative properties of multimodal speech.




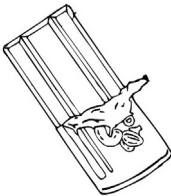
Acknowledgements

This research was supported by funding from NIH DC03172 to Dani Byrd, NSF 1551695 to Louis Goldstein, and the Frederick and Dorothy Quimby Memorial Scholarship to Samantha Danner. We would like to thank Sungbok Lee for his assistance with data collection. We would like to thank Drew Abney, Tessa Verhoef, and one additional anonymous reviewer, as well as guest editor Bodo Winter and the journal editor, Taehong Cho, all of whom gave very helpful comments on earlier versions of this manuscript. Finally, we are grateful to the late Eric Vatikiotis-Bateson for inspiring this work, for helping implement the software used for analysis, and for giving commentary on earlier versions of this work.

Appendix A. Experiment materials

Themes	Demonstration “Please demonstrate how you:”	Preference “State which option you prefer, and why”
Laundry	Fold your clean laundry 	1. Fold laundry or hang laundry on hangers? 2. Dry clean laundry or hand-wash laundry? 3. Dry laundry with a dryer sheet or without?

Appendix A (continued)

Themes	Demonstration "Please demonstrate how you:"	Preference "State which option you prefer, and why"
Jars & Containers	Open a tightly closed jar 	1. Screw-top jar or click-on lidded container? 2. Jarred or canned food?
Umbrellas	Open and use an umbrella 	1. Compact or full size umbrella? 2. Button closure or Velcro closure? 3. Umbrella or raincoat?
Bananas	Peel and eat a banana 	1. Not quite ripe banana or a little too ripe? 2. Sliced banana or whole banana? 3. Bananas, apples or oranges?
Candy	Unwrap a candy bar 	1. Do you like nuts in your chocolate bar or no nuts? 2. Foil wrapping or plastic wrapping? 3. Fruity-flavored candy or chocolate candy?

Appendix B. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.17632/d27w8kzrs2.5>.

References

Abner, N., Cooperrider, K., & Goldin-Meadow, S. (2015). Gesture for linguists: A handy primer. *Language and Linguistics Compass*, 9(11), 437–451.

Apple (2010). Quicktime Player 7.

Barbosa, A. V., & Vatikiotis-Bateson, E. (2013). FlowAnalyzer. Retrieved from <https://www.cefala.org/FlowAnalyzer/>.

Barbosa, A. V., Déchaine, R.-M., Vatikiotis-Bateson, E., & Yehia, H. C. (2012). Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. *The Journal of the Acoustical Society of America*, 131(3), 2162–2172.

Barbosa, A. V., Yehia, H. C., & Vatikiotis-Bateson, E. (2008). Linguistically valid movement behavior measured non-invasively. In R. Gucke, P. Lucey, & S. Lucey (Eds.), *Auditory visual speech processing* (pp. 173–177). Australia: Queensland.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2), 495–520.

Beattie, G., & Shovelton, H. (2000). Iconic hand gestures and the predictability of words in context in spontaneous speech. *British Journal of Psychology*, 91(4), 473–491.

Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. Retrieved from <http://www.praat.org/>

Bolinger, D. (1968). *Aspects of language*. New York: Harcourt, Brace & World Inc.

Busso, C., & Narayanan, S. (2006). Interplay between linguistic and affective goals in facial expression during emotional utterances. In 7th International Seminar on Speech Production (ISSP 2006) (pp. 549–556).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 335–359.

- Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In R. N. Campbell (Ed.), *Recent advances in the psychology of language: Formal and experimental approaches* (pp. 347–360). New York: Plenum Press.
- Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics*, 24(2), 209–244.
- Danner, S. G. (2017). *Effects of speech context on characteristics of manual gesture* (Unpublished doctoral dissertation). University of Southern California.
- Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Data for: Quantitative analysis of multimodal speech data, v5 [Dataset]. Mendeley Data. Retrieved from <https://doi.org/10.17632/d27w8kzrs2.5>.
- de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284–311). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511620850.018>.
- de Ruiter, J. P. (1998). *Gesture and speech production* (Doctoral Dissertation). MPI Series in Psycholinguistics. Radboud University Nijmegen.
- Duncan, S. (2005). Annotative practice (under perpetual revision). In *Gesture & thought* (p. Appendix). Chicago: University of Chicago press.
- Enfield, N. J., Kita, S., & de Ruiter, J. P. (2007). Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39(10), 1722–1741. <https://doi.org/10.1016/j.pragma.2007.03.001>.
- Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. In LREC: Workshop on multimodal corpora (Vol. 6, pp. 86–91). Malta.
- Friesen, W. V., Ekman, P., & Wallbott, H. (1979). Measuring hand movements. *Journal of Nonverbal Behavior*, 4(2), 97–112. <https://doi.org/10.1007/BF01006354>.
- Fuhrman, R. (2014). *Vocal effort and within-speaker coordination in speech production: Effects on postural control* (Master's Thesis). University of British Columbia.
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8–11.
- Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1–3), 185–203.
- Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, 22(3), 313–336. <https://doi.org/10.1080/01690960600632812>.
- Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica*, 32(C), 101–125.
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.
- Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, 8(1), 1–26.
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... Hunt, T. (2017). caret: Classification and regression training. Retrieved from: <https://cran.r-project.org/package=caret>.
- Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471.
- Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2), 179–214.
- Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, 36(8), 1404–1426.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science*, 15(2), 133–137.
- Munhall, K. G., Ostry, D. J., & Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance*, 11(4), 457–474.
- Namoodiripad, S., Lenzen, D., Lepic, R., & Verhoef, T. (2016). Measuring conventionalization in the manual modality. *Journal of Language Evolution*, 1(2), 109–118.
- Nobe, S. (2000). Where do most spontaneous representational gestures actually occur with respect to speech? In D. McNeill (Ed.), *Language and gesture* (pp. 186–198). Cambridge: Cambridge University Press.
- Ostry, D. J., Keller, E., & Parush, A. (1983). Similarities in the control of the speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(4), 622–636.
- Ozyurek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130296.
- Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, 42, 1–11.
- Parrell, B., Lee, S., & Byrd, D. (2013). Evaluation of prosodic juncture strength using functional data analysis. *Journal of Phonetics*, 41(6), 442–452.
- Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, 45(2), 329–343.
- R Core Team (2017). R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://doi.org/10.1007/978-3-540-74686-7>.
- Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4), 226–231.
- Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research*. Illinois: Evanston. Retrieved from <https://cran.r-project.org/package=psych>.
- Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J.-L. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech Language and Hearing Research*, 51(6), 1507–1521.
- Roustian, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *Speech prosody 2010-5th international conference on speech prosody* (p. 100110:1–4). Chicago.
- Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, 57, 283–300.
- Rusiewicz, H. L. (2011). Synchronization of prosodic stress and gesture: A dynamic systems perspective. In *Proceedings of GESPIN 2011*.
- Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4), 333–382.
- Schlangen, D. (2014). Sync your videos using reference audio. Retrieved June 13, 2016, from http://www.dsg-bielefeld.de/dsg_wp/wp-content/uploads/2014/10/video_syncing_fun.pdf.
- Shattuck-Hufnagel, S., Ren, P. L., & Tauscher, E. (2010). Are torso movements during speech timed with intonational phrases? In *Proceedings of speech prosody 2010* (pp. 2–5). Chicago.
- Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). In *Hand keypoint detection in single images using multiview bootstrapping* (pp. 4645–4653). IEEE.
- The MathWorks Inc (2013). *MATLAB*. Massachusetts: Natick.
- Turvey, M. T. (1990). Coordination. *American Psychologist*, 45(8), 938–953.
- Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. (2014). Articulatory coordination of two vocal tracts. *Journal of Phonetics*, 44, 167–181.
- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209–232.
- Westlund, J. K., D'Mello, S. K., & Olney, A. M. (2015). Motion tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PLoS ONE*, 10(6), e0130293.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, New York. Retrieved from <http://ggplot2.org>.
- Winfree, A. T. (2001). In J. E. M. L. Sirovich, S. Wiggins, (Eds.) *The geometry of biological time* (2nd ed., Vol. 12). New York, NY: Springer New York.
- Xiong, Y., & Quek, F. (2003). Gestural hand motion oscillation and symmetries for multimodal discourse: detection and analysis. In *2003 Conference on computer vision and pattern recognition workshop* (pp. 58). Madison, WI: IEEE.
- Xiong, Y., & Quek, F. (2006). Hand motion oscillatory gestures and multimodal discourse analysis. *International Journal of Human-Computer Interaction*, 21(3), 285–312.
- Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *From sound to sense* (pp. 97–102). Cambridge, MA.
- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2), 23–43.