COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12718

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12718

Complex Communication Dynamics: Exploring the Structure of an Academic Talk

Camila Alviar,[a,b] Rick Dale,[a,b] Alexia Galati[a,b,c]

[a]*Cognitive and Information Sciences, University of California, Merced*
[b]*Department of Communication, University of California, Los Angeles*
[c]*Department of Psychology, University of Cyprus*

Received 4 October 2017; received in revised form 25 January 2019; accepted 4 February 2019

Abstract

Communication is a multimodal phenomenon. The cognitive mechanisms supporting it are still understudied. We explored a natural dataset of academic lectures to determine how communication modalities are used and coordinated during the presentation of complex information. Using automated and semi-automated techniques, we extracted and analyzed, from the videos of 30 speakers, measures capturing the dynamics of their body movement, their slide change rate, and various aspects of their speech (speech rate, articulation rate, fundamental frequency, and intensity). There were consistent but statistically subtle patterns in the use of speech rate, articulation rate, intensity, and body motion across the presentation. Principal component analysis also revealed patterns of system-like covariation among modalities. These findings, although tentative, do suggest that the cognitive system is integrating body, slides, and speech in a coordinated manner during natural language use. Further research is needed to clarify the specific coordination patterns that occur between the different modalities.

Keywords: Multimodality; Communication; Dynamic complex systems; Situated cognition; Extended mind; Lecturing

1. Introduction

Human communication is highly multimodal. In fact, some researchers have described it as intrinsically multimodal (e.g., Enfield, 2013; McNeill, 1992). When talking to each other, humans do not only use words, but they also move hands and bodies, they vary pitch, they make longer or shorter pauses, they may rush or slow down, they may whisper or raise the voice, and they may even signal to objects in the environment when they

Correspondence should be sent to Camila Alviar, Cognitive and Information Sciences, University of California, Merced, 5200 Lake Rd., Merced, CA 95343. E-mail: malviarguzman@ucmerced.edu

become relevant to the conversation. All these different signals have been shown in previous research to carry important information that contributes to meaning in both production and comprehension (Vigliocco, Perniss, & Vinson, 2014). Indeed, in natural communication, the cognitive system coordinates all these behaviors simultaneously.

Yet the way in which humans control and coordinate multimodal signals during natural language use remains an underexplored puzzle. The present work takes a step towards filling important gaps in research on multimodality. First, as our review in the next section suggests, most research on multimodality focuses on the use of only a few modalities at a time, often in the context of very specific and controlled stimuli or tasks. Second, much of the research to date focuses on how language users vary a given modality when packaging information at the word or sentence levels. This scope overlooks the fact that natural human communication takes place in discourse contexts that demand extended, complex performances, beyond the sentential level. Third, most research on multimodality examines the coordination of different signals either by applying qualitative methodologies or using quantitative methodologies on highly controlled tasks.

In this work, we address these issues in the context of a specific complex performance: the delivery of an academic talk. We obtain automated and semi-automated multimodal measurements from naturalistic video data and apply quantitative methods to explore the patterns of dynamic organization and coordination among different modalities, over time. During lecturing, as in any other communicative task, speakers have to coordinate simultaneously different sources of information to successfully transmit their message. Speech, gestures, gaze, lecture slides, and so on, all convey information and need to be coordinated in a way that supports both clarity and engagement, and accommodates the memory and attentional constraints of the speaker and the audience at the same time (e.g., see Abrahams, 2016 for review of a "big data" approach to behaviors supporting successful presentation).

Our goal is to make an initial foray into the manner in which various multimodal signals change over the course of a complex performance, such as giving a talk. There is considerable relevant research that precedes our present undertaking, although much of it has focused on a single modality, or on how different modalities predict talk quality. In the next section, we review prior research on how different modalities—specifically, body movement, voice, and the use of visuals—are deployed during the course of a talk (Section 1.1). Then, we present some theoretical frameworks that accommodate the coordination of multimodal signals in extended communication, including situated cognition, the "extended mind" view, and dynamical systems (Section 1.2). Finally, we describe our methods in more detail and state our exploratory predictions (Section 1.3).

1.1. Adapting various modalities during a complex performance

1.1.1. Body movement

In the study of discourse and pragmatics, the way language users move their bodies has been thought to be central to their extended natural performance (Goodwin, 2000; Selting, 2010). Gestures, for example, are tightly intertwined with speech and

complement spoken information in a non-redundant way (McNeill, 2008). In many approaches—including cognitive linguistics, conversation analysis, and gesture psycholinguistics—speech-accompanying gestures have been studied as conceptually and structurally critical for discourse management (for review and analysis, see Wehling, 2018). Gestures tend to occur in temporal alignment with pitch accents and help group-related intonational phrases in bigger semantic units (McClave, 1994; Valbonesi et al., 2002; Yasinnik, Renwick, & Shattuck-Hufnagel, 2004). Postural changes, for their part, accompany shifts in conversational topics, as well as in conversational turns (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001).

Some studies have examined body movement specifically in the context of giving a talk. Although several nuanced patterns are reported, there is not yet a consensus about how body movement is implicated in such an extended performance. For example, Batrinca, Stratou, Shapiro, Morency, and Scherer (2013) found that short presentations were judged by experts to be most effective when speakers gestured and paced more. In political speeches, motion energy was also shown to correlate positively with persuasiveness and overall effectiveness (Scherer, Layher, Kane, Neumann, & Campbell, 2012), and the amount and type of movement was found to predict perception of diverse personality traits, such as agreeableness and extraversion (Koppensteiner & Grammer, 2010).

Much of the work on extended communication rarely examines *how* speakers organize their body over time across their complex performance, whether there are certain interparticipant patterns, and how much individual variability there is. We address these issues in the present work.

1.1.2. Voice

Speakers use various acoustic parameters to structure information during speech. Acoustic measures have been shown to relate to the information focus of the sentence. For example, in a series of studies manipulating informational focus in sentences (e.g., subject, verb, or object), Breen, Fedorenko, Wagner, and Gibson (2010) found that speakers naturally employ higher intensity, longer duration, and higher mean F0, to highlight the focal information. Intonational patterns also seem to narrow the search space of possible interpretations of an utterance (Wilson & Wharton, 2006). They help listeners predict later elements of the sentence and are specially used by speakers when an ambiguous interpretation of a sentence is possible (Snedeker & Trueswell, 2003). Prosodic boundaries signal higher meaning groupings across sentences that help listeners understand hierarchical dependencies in the information (Tseng, Pin, Lee, Wang, & Chen, 2005), and they reliably convey features of the syntactic structure of the message (Schafer, Speer, Warren, & White, 2000). Additionally, speech rate has been also found to vary depending on the complexity of the linguistic structures being used to convey a message (Cohen Priva, 2017), decreasing when less frequent words and structures are used and increasing for simpler linguistic constructions.

In much of this work, the acoustic feature of interest lies at the word or sentence levels. These are important levels of analysis, of course. In order to build cognitive theory, it is essential to understand the factors that guide the modulation of these signals at

the word or sentential levels (e.g., frequency, contextual predictability of lexical information, etc.), as well as the interplay of these factors. For example, Breen et al. (2010) showed that additional factors, such as the speaker's awareness of prosodic ambiguity in a given information structure, influence the use of prosodic features, with speakers producing different patterns for contrastively (vs. non-contrastively) focused elements.

In the present work, we wish to scale up the study of vocal features to broader units of analysis in extended communication. In our domain of interest (i.e., giving a talk) vocal features are in fact the signal most commonly examined. Various aggregate acoustic features such as F0 range and variability, speech rate, and disfluencies were found to be related to teacher effectiveness in short lectures (Schmidt, Andrews, & McCutcheon, 1998). Pitch variation and speech rate were also shown to relate to perceptions of liveliness during presentations (Hincks, 2005). In analysis of extended classroom discourse, with a focus on initiation-response-feedback exchanges, Hellermann (2003) found that different prosodic packaging was used for different types of feedback.

Some researchers have also analyzed speech patterns in other domains of natural language performance, such as political speeches or news stories. Work by Hirschberg and colleagues has shown several correlates between acoustic variables and higher level aspects of speeches, such as discourse structure (Grosz & Hirschberg, 1992; Hirschberg & Pierrehumbert, 1986) and a speaker's charisma (Rosenberg & Hirschberg, 2005). The standard deviation of acoustic measures, for example, predicts higher charisma in political speeches (Rosenberg & Hirschberg, 2005; see also Shim et al., 2015). Such cues may also serve as strong markers of genre or interactive style, along with other cues such as lexical features (Hirschberg, 2000; Jurafsky, Ranganath, & McFarland, 2009).

1.1.3. Visuals

PowerPoint slides have become the norm in scientific presentations. Despite this, few studies have analyzed the *dynamics* of talk delivery using slides. In some domains, such as educational research and communication, the use of slides and visuals has been a critical subject of study (for reviews, see Levasseur & Kanan Sawyer, 2006; Vyas & Sharma, 2014). Such research shows that fewer words on slides improve presentation effectiveness (Chen, Leong, Feng, & Lee, 2014), and that the dynamic elements of such visuals are an attractive feature (Moulton, Türkay, & Kosslyn, 2017).

The manner in which these slides are delivered, however, is rarely a focal point of analysis. Pacing and relative rhythm with vocal and bodily modalities are unexplored. Some survey-based methods asking presenters what they do in their talks suggests considerable individual variability (Brock & Joglekar, 2011).

1.1.4. Multimodal coordination of signals

It is important to note again that considerable prior research supports our position that these various signals would be integrated during extended communication. In the case of co-speech gesture, evidence for this integration is plentiful (McNeill, 2000, 2008). Gesture has been shown to complement spoken information in speakers' narrations (Cassell

& McNeill, 1991; Melinger & Levelt, 2004) and to aid listeners in comprehension (Cassell, McNeill, & McCullough, 1999).

In the context of teaching, Pozzer-Ardenghi and Roth (2004, 2007), using a qualitative approach, found that the gestures and body orientation of instructors during lecture conveyed meaning that amplified and supplemented what was being said by drawing attention to important details. This evidence was taken to be consistent with the view that words, gestures, and visual aids come together to form a holistic meaning unit during the communicative act. Recent quantitative evidence supports the notion that expressiveness during speech connects both prosody and body motion (Pouw & Dixon, 2018; Voigt, Podesva, & Jurafsky, 2014).

1.2. Theoretical accounts for multimodal coordination

In this section we review some theoretical accounts that we consider useful for characterizing the coordination of multimodal signals in extended discourse. As our review so far suggests, speakers coordinate an array of multimodal cues, including external tools, during lecturing and other discourse domains.

From a *situated cognition* perspective, the use of multimodal signals, slides included, can be beneficial for cognition. Using external representations when thinking facilitates complex thought by transforming mental computation into perception and action and leveraging the advantages of these two processes. External representations (the slides in this case) also serve as shareable objects of thought that offer a common and persistent referent for different people to attend to and reason with at the same time (Kirsh, 2010). The slides, then, can help the speaker and the audience reduce the cognitive load associated with conveying or understanding the message by offering a shared external representation that supports thinking.

From a more radical point of view, the slides can even be said to become a constitutive part of the speaker's and the audience's cognitive systems. Such a view is known as the *extended mind hypothesis*. According to this theory, first outlined by Clark and Chalmers (1998), the cognitive system and its processes are not only "internal" but also involve processes that extend outside the head into artifacts and the environment. In this view, the tools we use while thinking function as cognitive aids that, by sharing some of the representational load of the task at hand, become active elements of the cognitive system. As such, these tools need to be included in any explanatory account of that system.

Classical work within this theoretical framework has offered extensive descriptive evidence from complex tasks such as operating airplanes (Hutchins, 1995) or navigating the open sea without instruments (Hutchins, 1983). Researchers in this tradition argue that it is necessary to attend to the informational capabilities of the sociotechnical system as a whole (including person and tools), and not just to the cognitive abilities of the individual minds operating the machines.

In the case of interpersonal communication, Clark (2003, 2005) proposes that to understand meaning and communication we ought also to extend the window of analysis to include other material elements besides words. The gestures we make, the way we place

ourselves and the objects in the space with respect to other elements in the situation, and the specific moment at which we do it serve communicative functions of their own. Signaling techniques, such as pointing and placing, among others, facilitate coordination during interpersonal communication and ground the speakers' message to the material world. In line with this proposal, Hutchins and Palen (1997) showed that to understand how meaning is created and communicated within a system, it is necessary to attend to all the modalities involved. The authors found that aircrew members engaging in a problem-solving simulation task communicated complex messages by coordinating their words, their spatial orientation, and the organization of their gestures in relation to the tools in the environment. Each of the modalities—external tools included—participated in the creation of the message, which could only be fully understood when analyzing the different modalities in relation to one another.

Besides being multimodal, communication unfolds over time in an organized way. *Dynamical systems* approaches to the understanding of the mind offer novel methodologies to quantitatively characterize the dynamics of a system (i.e., the behavior of the system over time) and measure coordination between the elements that produce them (Richardson, Dale, & Marsh, 2014). These approaches conceptualize cognitive processes as emergent phenomena resulting from the multiplicative interaction of simpler interdependent elements and processes (McClelland et al., 2010). The cumulative effects of the interactions between elements creates patterns of organization that can be measured at different timescales of the system's behavior (Kello et al., 2010). Self-organizing systems also tend to exhibit "synergies" between the elements that comprise them, such that multiple parts of the system come to operate together as a functional whole, exhibiting patterns of coordination and compensation that respond closely to the environmental constraints and the task demands (Kelso, 2009; Riley, Richardson, Shockley, & Ramenzoni, 2011; Shockley, Richardson, & Dale, 2009). In dialogue, interactional routines like turn taking, for example, seem to emerge from the low-level information of multiple channels. The words and syntactic structures being used (de Ruiter, Mitterer, & Enfield, 2006), the prosodic boundaries of intonational phrases (Bögels & Torreira, 2015), postural changes (Cassell et al., 2001), and the patterns of gaze orientation (Rossano, 2013) all seem to play a role on the coordinated dance of turn taking. Speech rate also seems to help with turn-ending prediction, as it may mediate the synchronization of conversational pace between partners (Wilson & Wilson, 2005).

Examining the way behaviors are temporally organized, independently and in relation to each other, offers a window into the characteristics of the system that is producing them. Studying multimodal communication within this framework allows us to explore the structure and coordination patterns of the cognitive system during communication in general, and during the presentation of complex information, in particular. Additionally, this approach allows us to expand on the situated cognition literature by making the first steps toward quantitatively exploring the way in which an extended cognitive system coordinates its parts. If extended cognitive systems do in fact emerge from the interaction of a person and her tools, resulting in interdependent systems, we should find evidence of

structure in those systems' behavior across different time scales, and patterns of coordination between their behaviors.

1.3. The present study

In this work, we use a natural dataset of academic lectures to conduct an exploratory analysis of the temporal structure of multimodal behaviors during this complex extended performance. Analytically, we use a combination of automated and semi-automated methods. As our review suggests, the extant literature typically examines the coordination of modalities either by applying qualitative methodologies (e.g., Hutchins & Palen, 1997) or using quantitative methodologies on highly controlled tasks (e.g., Valbonesi et al., 2002). Some exceptions include work that has used motion tracking or computer vision object-tracking to capture various features of the speakers' body movement (Chen et al., 2014), and work that has used automation to assess public speaking skills (Batrinca et al., 2013; Scherer et al., 2012) and to explore multimodal coordination in short sentences (Voigt et al., 2014). But even so, there remains a paucity of research on the structure of these signals as they vary in time.

Automated and semi-automated methods are useful tools for studying extended communication. Uncovering organizational patterns during communication requires large amounts of data that are costly to obtain using the traditional hand-coding methods common to psychology. The use of extensive videos and automated techniques for analyzing them allows us to frequently sample behaviors of interest with reduced effort and minimal time costs, while still achieving sufficient data quality. Frame-differencing methods are a useful application of computer vision to measure the movement in a video by quantifying the amount of pixel change across consecutive video-frames (see Paxton & Dale, 2013 for a review). When applied to videos in which the observed movement is coming from an empirically relevant source, or in which this source can be easily isolated, these methods offer a good proxy for variables of psychological interest, such as body movement and coarse gestures (see Pouw, Trujillo, & Dixon, 2018 for a comparison with motion tracking methods). Similarly, software like Praat (Boersma & Weenink, 2010) offers a broad set of tools to analyze and easily quantify diverse acoustic properties of audio signals, such as the intensity, fundamental frequency, formants, and power spectrum. This allows rapid access to prosodic information that contains clues to interesting psychological components of discourse.

In the present work, we are interested in addressing two main questions: First, how do body, slide transitions, and prosodic components of speech—such as pitch, rate, and amplitude—change during the development of a talk as time advances and constraints change? And secondly, how are the changes in these different modalities dynamically related to each other? Are there covariation patterns between them? We focus on a segmentation that is relatively coarse-grained—analyzing subsets of behavior between 4 and 8 min in duration. The purpose is to find if particular periods of time in the extended performance enjoy a kind of highlighting in the multimodal signals we analyze.

Motivated by our review above, there are a number of possible patterns which we might observe here. Though we take an exploratory approach, it is instructive to consider these possibilities. One possibility is that the modalities show a linear *increase* over time: For example, speakers could show more irregular movements as their social motivations change during the talk (Koppensteiner & Grammer, 2010), increasing their overall body movement. Similarly, speakers could increase their slide rate towards the end of their talks as they run out of time to cover the material. Conversely, a second possibility is that the modalities show a linear *decrease* over time: Speakers attenuate their gestures (Galati & Brennan, 2014; Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015) and their articulation (Galati & Brennan, 2010) as they build common ground with their audience. Such adaptation would be consistent with a *lowering* of voice signals and body movement across a presentation. Speech rate may also decrease over time as people get into more complex parts of their talks that require more complex linguistic constructions (Cohen Priva, 2017). A third possibility is that the modalities show a *U-shaped* pattern: If the middle of the talk carried most of the new information, speakers might utilize increased vocal and bodily expressiveness (Galati & Brennan, 2010) to elaborate the novel concepts at this point of the talk, leading to an inverted U pattern. All these described patterns suggest that some kind of packaging may be dependent upon the *region of time* during an extended performance. An additional possibility, however, is that we fail to find any structure over time in these signals. Such a null effect might suggest that *local* effects of prosodic or bodily variation are sufficient to explain performance within this circumscribed discourse context. In that case, variation may be detectable only at these lower levels (e.g., when considering words or sentences).

Beyond considering each modality separately, we are also interested in the covariation of these multimodal signals to tap into their interdependence and synergies as components of the same system. We do not posit any hypotheses in advance, as there are several possibilities. For example, voice and body movement may pattern in a parallel, reinforcing manner (e.g., Voigt et al., 2014), or it may instead pattern in a more complementary fashion. The use of slides may also bear different relationships with the other signals—for example, an increase in the use of slides may be associated with a decrease in some signal (e.g., reduced body movement) but increase in another (e.g., faster speech rate). We examine the various possibilities in the systematic covariation of these multimodal signals by assessing their "compressibility" through principal component analysis (PCA).

2. Method

2.1. Data

The video recordings of 30 lectures given in the Cognitive and Information Sciences seminar series at UC Merced were analyzed. This seminar series covers a broad range of topics in Cognitive Science, and recordings of its talks are publicly available in YouTube[1] and Vimeo.[2] Each talk is typically an hour long followed by a 30-min (on average)

Q&A session, which was not included in the recording. Each lecture was videotaped from a fixed point within the audience, with the camera set up on a tripod. The camera was usually let to run without any modifications to its focus or position for the length of the talk. This made the videos suitable for meaningful analysis with frame-differencing techniques described in the next section.

The recordings included in this study were selected using the following criteria to permit automated video and audio analyses: (a) the camera was fixed in the same position and focus during the whole lecture, (b) the speaker and the slides were always visible in the video, (c) the heads of audience members did not occlude the speaker from the camera view at any point, (d) there was low overlap between the speaker and the slides (determined through visual inspection), and (e) the quality of the audio track was sufficient (i.e., low background noise, no echo) to perform automated analysis in the speech signals.

The resulting sample of 30 videos used in the study included talks by 8 females and 22 males, from various fields, including philosophy, psychology, neuroscience, linguistics, computer science, and math. The sample also involved participants from diverse locations within the United States: Of the 30 speakers, 15 were affiliated with institutions on the West Coast of the United States, and 15 with institutions in other parts of the country. The presentations included in the analysis were between 34.6 and 81.3 min long, with a mean duration of 56.45 min (SD = 10.21 min).

2.2. Instruments and procedure

2.2.1. Video analysis

The videos were downloaded from YouTube and Vimeo in the best possible quality using the iSkysoft Video Downloader (i.e., 1080p for YouTube videos, and 360p for Vimeo videos). These were then converted to AVI format using the Any Video Converter software. The Optical Flow Analyzer (Barbosa, Yehia, & Vatikiotis-Bateson, 2008) was used to conduct the frame-by-frame analysis. This software compares the locations of the pixels in the current frame with the location of the pixels in the previous frame of the video. It calculates the magnitude and direction of the displacement of every pixel within some defined area of interest and, using the frame sampling rate, produces a velocity vector for each of the pixels in that area. The Optical Flow Analyzer then sums all the individual vectors coming from each pixel in the delimited area and returns a vector reflecting the overall pixel change within each area of interest for every frame of the video (25 frames/s). Two main areas of interest were defined for each video (see Fig. 1): one demarcating the area in which the speaker was moving during most of the lecture; and the other one demarcating the area occupied by the slides. The areas of interest were placed in a way that avoided any overlap between them to make sure that any pixel change detected in any of them, presumably corresponded only to the movement of the object they were enclosing (i.e., either the speaker or the slides).

We faced two general issues in using the Flow Analyzer technique with these videos. First, occasionally, the speaker would step in front of his or her slides. We carefully

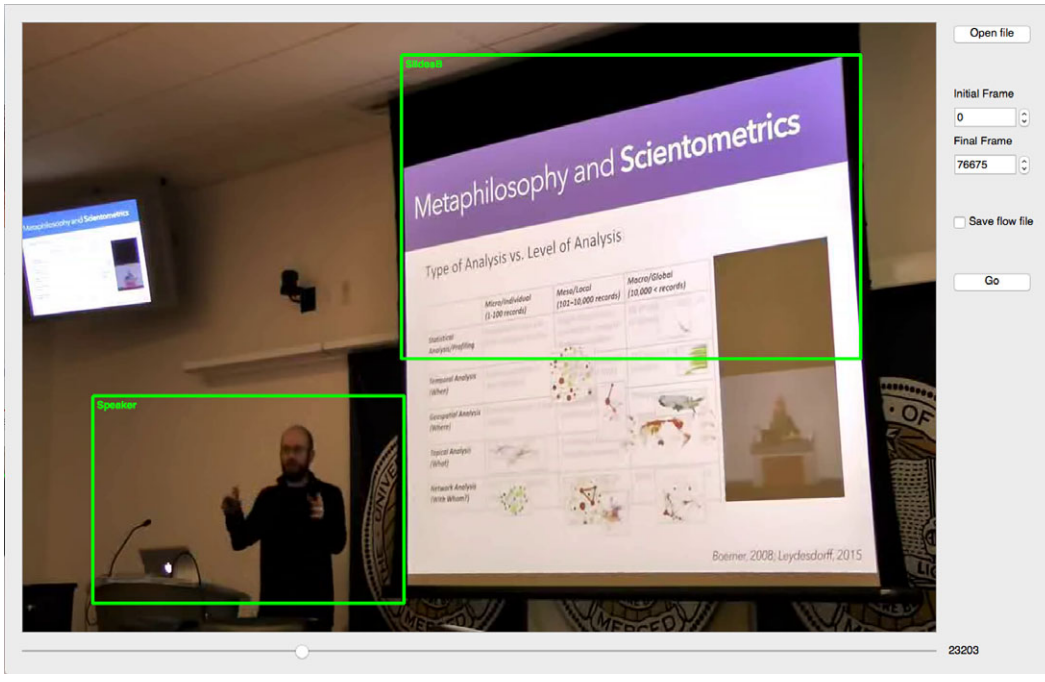

Fig. 1. Screenshot of the Optical Flow Analyzer showing the placement of the two areas of interest in one of the videos. Note that the areas are placed in a way that avoids information from the slides being captured within the speaker's area, as well as movement from the speaker being captured within the slides' area.

selected videos which minimized this artifact, though there remained some of these instances in the videos. However, this occasional overlap was neither captured in the speaker's nor the slides' movement data: The areas of interest were placed so that they did not include regions where there was frequent overlap between the speaker's movement and the slides (see Fig. 1). Therefore, the area corresponding to the slides usually included only the top portion of the slides to avoid capturing any of the speaker's movement (e.g., when walking in front of the slides) in the slides' area of interest. This strategy allowed us to avoid erroneously registering the speaker's movement as slide movement and vice versa. However, it also resulted in data loss from the lower parts of the slides. This, as discussed below, makes our analysis less sensitive to slide animations (e.g., the use of animated bullet points) in comparison to slide transitions (i.e., changes from one slide to the next). Second, slide changes sometimes changed the illumination of the presenter's body, which could in turn briefly cause a change in pixels, leading to a punctate moment of spurious body movement. In order to ensure that the algorithm was producing coherent results, untainted by these artifacts, we carried out a validation analysis that is presented in a later section.

For each video, we divided each time series obtained from the Optical Flow Analyzer into 10 windows of equal size (each window representing 10% of the duration of the

presentation in question—the duration of the window varied depending on the length of the talk) and got the aggregated measures of the speaker movement, the slide changes, and the prosodic components of the speech signal for each of the windows. This aggregation helps overcome the noise that may be present at finer-grained timescales: If subtle fluctuations in pixel change occurred from the slides, or from the speaker moving in front of the slides, these would only count as brief noise when averaging over several *minutes* of the presentation. Moreover, segmenting the signals into a number of windows as opposed to a number of minutes allowed us to align and compare pragmatically similar discourse segments of the talks (e.g., introduction to the topic, presentation of research findings, concluding remarks) regardless of the talk's exact duration. Two variables were obtained from the video analysis using R (R Core Team, 2018): body movement and slide change.[3]

2.2.1.1. Body movement: We obtained the mean pixel change within the speaker's area of interest for each window of each talk.

2.2.1.2. Slide change rate: As the slides change discretely, they produce big spikes of pixel change observed in the output from the Optical Flow Analyzer. We used an automated algorithm to find the spikes of pixel change in the slides signal. For this, we defined a threshold for each talk and identified the times in which big changes occurred. A minimum of 10 s between spikes was set to avoid counting multiple spikes coming from videos in the presentation, or other noise sources to be counted as slide changes. To determine the threshold, we plotted the standardized signal and determined visually the number of *SD*s required to minimize the amount of noise (i.e., constant small amounts of pixel change) and maximize the number of spikes (i.e., sporadic big amounts of pixel change) being captured. To validate the algorithm, we hand coded the slide changes (attending also to animations within the same slide) of three talks and compared these time series to the ones obtained using the algorithm. We calculated two common signal processing measures to evaluate the algorithm's performance: precision, which captures the ratio of correctly identified events versus the total number of events found by the algorithm; and recall, which captures the ratio of correctly identified events versus the total number of events actually present in the signal. When counting slide transitions (i.e., full slide changes) and slide animations (i.e., small changes within the same slide) as a slide change, the average precision and recall rates were 0.849 and 0.421, respectively. When considering only slide transitions as a slide change, the average precision and recall rates were 0.829 and 0.662, respectively, suggesting that the algorithm was especially successful at capturing full changes in the slides. For our analyses, we used the event count within each time window determined by the algorithm and computed the rate of slide change per minute.

2.2.1.3. Validation of video analysis: To validate the data obtained with the Optical Flow Analyzer, we selected and hand coded 200 5-s segments of five randomly selected videos (40 segments from each). The segments from each video were chosen

using an algorithm in R that identified and selected 10 5-s high body movement windows (i.e., windows with high mean and low *SD* pixel change from the speaker area), 10 5-s low body movement windows (i.e., windows with low mean and low *SD* pixel change from the speaker area), 10 5-s high slide change windows (i.e., windows with high mean and high *SD* pixel change from the slides area), and 10 5-s low slide change windows (i.e., windows with low mean and low *SD* pixel change from the slides area).

All three co-authors coded each of these segments for amount of body movement (on a scale from 1-*low movement* to 7-*high movement*) and presence of slide changes (on a binary 1-*Change*, 0-*No change* scale). The Krippendorff's alpha (Hayes & Krippendorff, 2007; R package: Gamer, Lemon, & Fellows, 2012) indicates there was good interrater reliability for both the body movement judgments ($\alpha = 0.768$) and the slide change judgments ($\alpha = 0.964$). To dichotomize the body movement variable (making it comparable to the information we used to select the segments: high or low movement), we calculated the mean movement out of the three raters' scores and performed a median split on the resulting values. Everything over and equal to the median movement rating was considered high movement, and everything below, as low movement. The comparison of the software's classification against the human raters' classification showed that (a) 46 of the 50 high body movement segments were judged to contain high speaker movement, (b) 47 of the 50 low body movement segments were judged to contain low speaker movement, (c) 41 of the 50 high slide change segments were judged to contain a slide change, and (d) 48 of the 50 low slide change segments were judged to contain no slide changes.

2.2.2. Audio analysis

The speech signals were extracted from the videos, filtered, and segmented using Audacity. The noise reduction feature with the default settings was used to reduce the background noise in the sound waves, and the regular interval labeling feature was used to segment the files in 10 parts. Praat (Boersma & Weenink, 2010) was used to calculate the fundamental frequency, speech rate, articulation rate, and intensity of every speech excerpt.

2.2.2.1. Fundamental frequency: The fundamental frequency was computed using the autocorrelation method ("To pitch" feature) in Praat. The pitch floor and ceiling were set to 75 Hz and 400 Hz, respectively, for male speakers, and to 100 Hz and 500 Hz for female speakers, following the recommendations in the Praat manual. The male ceiling was set 100 Hz higher than recommended because a good number of the male speakers used perceptually higher pitches when talking. The mean F0 for every time window was obtained. This measure was preferred over the range of F0 because the automated algorithm in Praat produces occasional spurious overestimations and underestimations of the pitch. While these occasional errors in the estimation have serious effects on the estimation of the range, they are less relevant for the calculation of the mean, making the latter a better measure.

2.2.2.2. Speech and articulation rates: The speech and articulation rates were calculated using De Jong and Wempe's (2009) Praat script. This script uses the peaks in the intensity contour to identify syllable nuclei. Speech rate is calculated as the number of syllables divided by the total duration of the signal; and articulation rate as the number of syllables divided by the duration of the voiced parts of the signal. The silence threshold was set to −26 dB, the minimum pause duration to 3 s, and the minimum dip between intensity peaks to 4 dB, given that we are working with noisy signals.

To validate the data, we randomly selected 100 30-s segments from five of the videos (20 segments each) and counted the number of syllables on each excerpt. The correlation between our syllable count and the algorithm's was 0.63.

2.2.2.3. Intensity: We used the "To intensity" feature of the Praat software to determine the intensity contours. The default settings were utilized. The mean intensity for each time window was calculated using the "energy" averaging method.

2.3. Data analysis

We conducted two separate analyses on these multimodal signals. First, we tested whether speakers consistently vary these behaviors across their talk. This first analysis tested each modality separately, independently of the others. We were interested in uncovering whether the cognitive system responds differently in various moments of the presentation as it advances. For example, do speakers begin with rapid speech and slow slides changes? And do they end on slower articulation and quick sequences of slide changes as time limitations become more pressing? Because we were interested in both linear and nonlinear patterns over time (see below), we built mixed-effects models for each of the dependent variables (pitch, body movement, etc.) using time window (1–10) both as a linear and a quadratic predictor. To control for the effect of between-subject variation in the observed trends, we introduced speaker identity as a random effect. A maximal random effect structure, with a random intercept and slope for speaker identity, was specified in the model following the recommendations of Barr, Levy, Scheepers, and Tily (2013) and Mirman (2014). The models were built in R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015), and their p-values were calculated using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

The second analysis, as foreshadowed in the introduction, explored the covariation *among* these behavioral, multimodal signals in order to capture their interdependence. If the cognitive system is coordinating these signals together, as interdependent parts, we should be able to substantiate these interactions through a quantitative method. As a first step in this direction, we used PCA, a dimensionality reduction technique, to explore if these multimodal measures are correlated and how "compressible" they are.

Though we used PCA, there are many dimensionality reduction techniques to choose from (Van der Maaten, Postma, & Van Den Herik, 2009). Given the relatively sparse data here (300 rows) and our exploratory purpose, we chose PCA as it is among the simplest

—it simply performs orthogonal rotation of our observed data. We used the prcomp (R Core Team, 2018) function in R, from the core "stats" package. For our purposes, we took the unrotated output of the prcomp function. Importantly, rotation would have no effect on the observed compression of our variables, only their interpretation. We return to this issue in the General Discussion, where we discuss future directions for dimensionality reduction.

The prcomp function performs PCA using singular value decomposition. Variables were neither centered nor scaled beyond our own transformations (as described below). We interpret the variance accounted for by a given component as its corresponding eigenvalue.

Our primary goal is to determine whether there is a compression among the modalities that we measured. In this PCA approach here, this would appear as a smaller number of components accounting for a disproportionate percentage of the variance in these data. There are many approaches to determining this in PCA, such as a bend in a scree plot of eigenvalues or cumulative variance with a cut-off. Here we report eigenvalues above a value of 1.0, indicating that a component is accounting for more variance relative to the original standardized data. We also report percentage of variance accounted for, expecting that a disproportionate variance will be accounted for by fewer components than the dimensionality of the observed data.

3. Results

3.1. Patterns of variation of the modalities over time

The first goal of this study was to explore and describe the ways in which the different modalities are organized and change as the presentation progresses and the constraints the speaker has to face change. Fig. 2 shows the average patterns of variation over time for each dependent measure. Most variables, with the exception of slide rate, exhibit a decreasing trend over time. In the case of body movement (Fig. 2, top left), speech rate (Fig. 2, middle left), and intensity (Fig. 2, bottom left), this decrement appears to be principally linear. However, in the last segment of the intensity time series (Fig. 2, bottom left) a sudden drop is visible. In the time series for articulation rate (Fig. 2, middle right) an inverted U-shaped pattern is observed. These two characteristics of the observed patterns suggest that nonlinear components could also be necessary to capture the distribution of the data.

Table 1 shows the raw and standardized coefficients for the mixed regression models of each dependent variable as a linear or quadratic function of time. We used linear mixed effects models with maximal nested random effects to help avoid false positives for each modality, by factoring in a more complex nested model of each individual speaker. Despite that, the fact that we tested multiple models—for the several modalities–does leave the possibility of Type I error lingering. The overall purpose of this paper was to explore these patterns in time, and so future work should follow up by expanding the available dataset in this domain and others, and perhaps expand model complexity (e.g.,

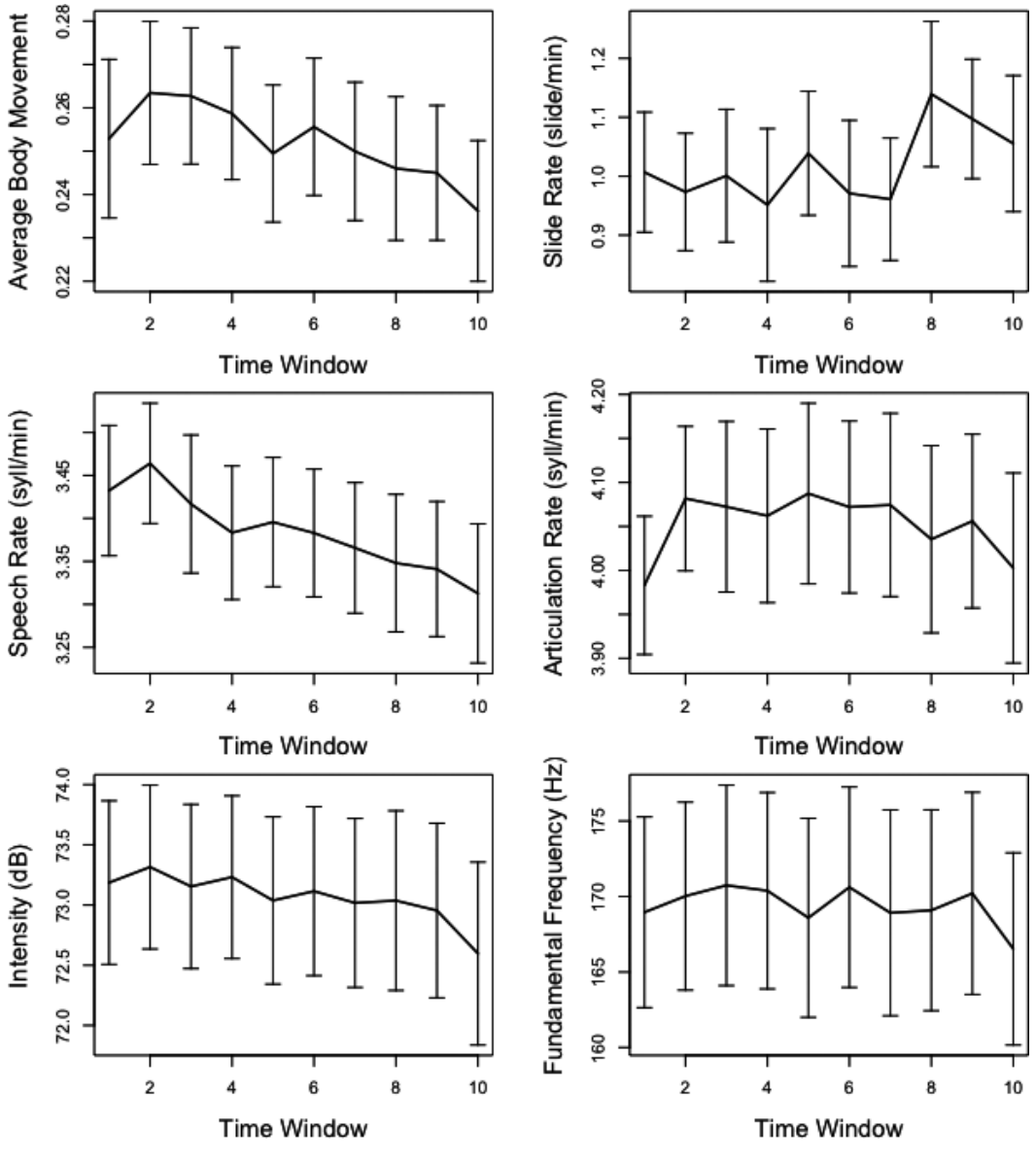

Fig. 2. Overall patterns of variation over time for the different modalities of information: body movement (top left), slide change rate (top right), speech rate (middle left), articulation rate (middle right), intensity (bottom left), and fundamental frequency (bottom right). The time windows correspond to equally long segments of the talks, each comprising 10% of the duration of them. Error bars indicate *SE* of the mean and are corrected by within-subject measurements.

including more nonlinear terms, individual differences, etc.). This may help fine-tune which of our observed effects are real, and which effects we may have missed (Type II error).

Table 1
Coefficients for the linear and quadratic mixed-effects models

DV(units)	Predictor	*B*	β	*SE*	*t*	*p*
Body movement (pixel change)	*t*	−0.0023	−0.026	0.013	−2.00	.054
	t^2	−0.00039	−0.0045	0.0031	−1.41	.17
Slide rate (slides/min)	*t*	0.012	0.019	0.017	1.13	.27
	t^2	0.0023	0.0037	0.0070	0.53	.59
Speech rate (syll/s total)	*t*	−0.014	−0.034	0.013	−2.57	**.016**
	t^2	−0.00017	−0.00042	0.0037	−0.11	.91
Articulation rate (syll/s voiced)	*t*	−0.0010	−0.0019	0.013	−0.14	.89
	t^2	−0.0036	−0.0069	0.0030	−2.26	**.032**
F0 (Hz)	*t*	−0.19	−0.0053	0.0067	−0.80	.43
	t^2	−0.080	−0.0023	0.0018	−1.24	.22
Intensity (dB)	*t*	−0.054	−0.014	0.0056	−2.55	**.016**
	t^2	−0.0087	−0.0023	0.00096	−2.40	**.023**

Note. The reported *SE*s correspond to the models using the standardized data. $p < .05$ are shown in bold.

The *t*-tests confirm our observations above, indicating a statistically significant linear effect of time on speech rate ($p = .016$) and intensity ($p = .016$), and a marginally significant effect ($p = .054$) on body movement. The negative sign of the slope for all variables suggests that the rate and intensity of speech, and the amount of movement decrease as the presentation advances (see Fig. 2: middle left, bottom left, and top left).

Significant effects for the quadratic component of the regression models were found for intensity ($p = .023$) and articulation rate ($p = .032$). In the case of intensity, this suggests that the decrease in the amplitude of the signal over time, observed in the bottom left panel of Fig. 2, is not completely linear: Intensity decreases at a somewhat regular pace, and it dramatically drops off at the last window. In the case of articulation rate, the significant quadratic term suggests that the number of syllables pronounced during actual voiced time increases as speakers go into the middle section of the talk and decreases again as they move into the final segments (see Fig. 2, middle right).

The effects of time on the different variables are modest, as observed by the small beta coefficients in Table 1. This suggests that the specific patterns we have examined may be insufficient to fully describe the ways in which the different modalities change during these extended periods of time. More specifically, individual differences may play an important role in the observed trends. In fact, an initial exploration shows that different individuals do seem to be quite variable in their organization of the modalities. For example, speaker 6 and speaker 29, whose body movement and speech rate data are presented in Fig. 3 (left and right, respectively), show completely opposite patterns for both of these modalities. Speaker 6's body movement decreases and her speech rate increases as the lecture advances, whereas speaker 29's body movement increases and her speech rate decreases over time. The high variability of the individual speakers' patterns may be causing the random effect structure in the model to take up most of the variance, which in turn may be reducing the contribution of time as a fixed effect.

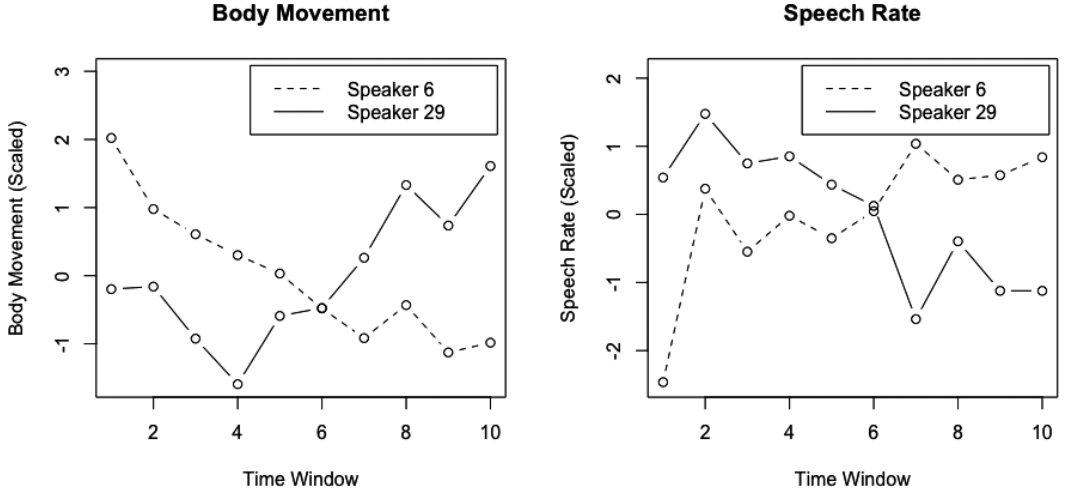

Fig. 3. Body movement (left) and speech rate (right) data for speaker 6 and speaker 29.The subjects show opposite patterns for each of the variables and exemplify the types of unique strategies that different speakers exhibit during their talks. Individual differences in the use of each modality might have a role in explaining the small effect of time in the mixed-effects models.

3.2. Covariation between modalities

The second goal of this study was to determine whether there were "system-like" properties in the speakers' multimodal behaviors—that is, to test if the different modalities varied in a coordinated way. We combined all the behavioral measurements, window-by-window, in one data matrix. Each of 300 rows in this matrix represented a speaker ($N = 30$) in a given 1/10th window, with each column reflecting one of the six modalities, as measured above. We used PCA to examine whether a small number of dimensions (components), discovered through the spectral decomposition of the data matrix, would adequately describe multimodal performance in a "compressible" manner. We applied PCA to two different covariation matrices: one standardizing the data across the whole sample, and the other standardizing the data within subject.

The standardization across the whole sample highlights the differences *between speakers* by scoring their performance on different variables relative to the group mean; thus, each speaker's score on a given variable would be high or low relative to the other speakers. This standardization procedure allowed us to ask questions about the covariation patterns in the distribution of "high" and "low" scores across the different variables. For example, do people who move more tend to have more slides and speak faster? Or, in contrast, do those who move more have fewer slides?

Table 2 shows the loadings of this PCA solution, as well as the eigenvalues for each of the factors. The first three components account for approximately 75% of the variance, have eigenvalues >1 (i.e., accounting for more variance than any single original variable), and suggest the existence of patterns in the strategies that different speakers adopt when

giving a talk. The first component accounts for 34.63% of the variance and involves a mixture of four variables: body movement, speech rate, articulation rate, and intensity. All these variables load positively on this component (they all have the same sign), suggesting that speakers who tend to speak faster also tend to speak louder and move more. The second component accounts for 20.94% of the variance and contains the variation of three variables: body movement, slide rate, and intensity. The signs of the loadings suggest that speakers who have more slide changes also tend to move more and speak more softly. The third component accounts for 17.33% of the variance and combines the variation of two variables: fundamental frequency and intensity. The signs of the loadings suggest that speakers who have a higher pitch tend to speak more softly.[4]

In a second exploration, we wished to investigate how modalities vary across the talk *irrespective* of the overall magnitude of the behavioral variables. In other words, if a speaker increases *her* body movement, does she also increase *her* slide changes? To explore this, we conducted a PCA on *z*-scores computed *within* speakers. In this standardization procedure, we used each speaker's individual mean for each variable to standardize their score for each of the time windows of that variable. This way of standardizing highlights the differences *between time windows*, allowing us to ask questions about the way the different variables co-vary at different moments in time as they go above or below the speaker's mean.

Table 3 shows the loadings and eigenvalues of this PCA solution. Only the first component has an eigenvalue >1, and it accounts for 29% of the variance. This solution is more difficult to interpret than the first PCA solution, as it shows less compressibility of the measures, as evidenced by the low percentage of variance accounted by the only component with an eigenvalue >1. Nevertheless, we attempt to unpack what the first component's loadings suggest. The first component (29% of the variance) mainly includes variance from speech rate, intensity, articulation rate, and F0. This suggests that as speakers increase their speech rate, they also tend to increase their pitch and their loudness.

In general, the covariation patterns observed between the modalities suggest there could be system-like synergies in how different modalities vary over time. However, it is difficult to uncover their specific patterns of covariation, since some of the dependent

Table 2
Principal component analysis solution standardizing across the sample

DV	C1	C2	C3	C4	C5	C6
Body movement	**0.22**	**0.66**	−0.01	0.72	−0.05	−0.01
Slide rate	−0.07	**0.74**	0.09	−0.64	0.15	0.07
Speech rate	**0.61**	−0.08	−0.09	−0.12	−0.15	0.76
Articulation rate	**0.57**	0.02	−0.03	−0.23	−0.50	−0.60
F0	0.06	0.04	**−0.95**	−0.05	0.28	−0.12
Intensity	**0.49**	**−0.12**	**0.28**	0.01	0.79	−0.22
Eigenvalues	**2.07**	**1.25**	**1.03**	0.68	0.62	0.32
Percent σ^2	**34.6**	**20.9**	**17.3**	11.4	10.4	5.3

Bold indicates the eigenvalues of the components that are greater than 1, their explained variance, and the loadings of those components that are greater than 0.1.

Table 3
Principal component analysis solution standardizing within subject

DV	C1	C2	C3	C4	C5	C6
Body movement	**−0.10**	0.79	−0.22	−0.08	−0.54	0.10
Slide rate	−0.09	0.31	0.93	0.17	0.07	0.00
Speech rate	**−0.53**	−0.26	−0.04	0.51	−0.24	0.58
Articulation rate	**−0.50**	−0.33	0.16	−0.34	−0.50	−0.50
F0	**−0.46**	0.12	0.00	−0.66	0.43	0.39
Intensity	**−0.49**	0.28	−0.25	0.39	0.46	−0.50
Eigenvalues	**1.74**	0.99	0.89	0.79	0.70	0.30
Percent σ^2	**29.0**	16.6	14.8	13.2	11.7	5.0

Bold indicates the eigenvalues of the components that are greater than 1, their explained variance, and the loadings of those components that are greater than 0.1.

variables load together in more than one component but exhibit contradictory relationships across components (see for example in Table 3 the loadings of body movement and articulation rate in C1 and C2).

Using orthogonal data rotation yielded only subtle features of compression. Probing further into the results of this PCA solution may require a larger sample size. Future explorations of the data may involve using more fine-grained time windows or including other relevant variables for the behavior under study (e.g., informational complexity of the information being conveyed). In addition, it may be useful to explore other dimensional reduction techniques (Van der Maaten et al., 2009).

4. Discussion

This study explored how different modalities of communication are used by speakers during the delivery of an academic talk, and the relationships between those modalities. In regards to our first objective—to explore how different modalities are organized and change over the course of the lecture—our findings suggest some small but significant effects. Of the 12 coefficients we tested (2 fixed effects per 6 modalities), four showed significant but small effects.[5] In line with a dynamical systems account, time predicted significantly the speakers' speech rate, articulation rate, intensity, and marginally the speakers' body movement. As the presentation advanced, speakers reduced their volume, their speaking rate, and their movement. They also had higher articulation rates in the middle of the talk than at the beginning.

A simple and quite obvious explanation for this general decrease across modalities is a fatigue effect, given the speakers' sustained performance over the 1-hr lecture. However, another, perhaps more interesting possibility is that the observed trends might be related to an increase in informational complexity as the talk advances. As more complex or less established information is reached (e.g., the results or conclusions of the presented research), it is reasonable to expect presenters to speak more slowly, make longer pauses,

and decrease their volume as a result of uncertainty or difficulty in conveying the information. Consistent with this possibility, speech rate has been shown to depend on informational complexity, decreasing as complexity increases (Cohen Priva, 2017). Testing this possibility would require further exploration of the data. Future analyses of datasets of this kind may involve classifying or quantifying the slides' content in order to gauge the complexity of the information presented at each time window.

The magnitude of the effect sizes resulting from the regression analysis suggests that the dependency of the different modalities on time is limited. As we have mentioned before, this could reflect the importance of individual differences. It is possible that speakers vary widely in their use of the different modalities during the talk, such that the unique impact of time on each speaker (captured by the maximal random effect structure) accounts for most of the variance in the models, obfuscating the impact of time on individual modalities. Some initial explorations of the individual patterns of the speakers seem to support this possibility, as illustrated in Fig. 3.

We also explored model comparisons between models with different random effect structures to select those with the best fit, following a reviewer's suggestion. Ultimately, we decided to continue reporting the models with the maximal random effect structure here for two reasons. First, specifying the maximal random effect structure is a more conservative approach that reduces the possibility of Type I error (Barr et al., 2013). Second, in light of the individual differences we have noted, we wished to control for the effects of individual subjects on the general trend (thus maintaining the maximal random effect structure). Model selection could be most valuable when having to specify more complex models, with additional parameters. In future work, such models could include the slide content or informational complexity, or the state of the other modalities at current or previous time steps. The best fitting of these models could provide insight about the specific variables that predict the organization and change of communication modalities during lectures. Doing model selection for the parameters of the random effect structure could also be a good way of testing for individual differences: The change in R-squared and in model fit indices between a maximal model and one without a random slope could illustrate the contribution of individual trends to the change of modalities over time.

We analyzed relatively large bin sizes across time, averaging measurements into 10 temporal bins. This methodological choice enabled us to stabilize behavioral measurements (e.g., reduce noise from small fluctuations in pixel change by averaging over several minutes) and to compare pragmatically similar discourse segments across speakers (e.g., the dip in many of the behavioral signals at the conclusion of the talk, as illustrated in Fig. 2, or a peak in slide change rate in window 8, top-right panel of Fig. 2).[6] However, our choice of relatively long time windows may have contributed to the modest effect sizes observed, since constraints at longer time scales (e.g., at the approximately 6-min time windows we have used) may be less important than local constraints in guiding the organization of behavior during a talk. As we have discussed, it is possible that the way in which speakers recruit multiple modalities during a talk is significantly influenced by the complexity of the information presented and, specifically, by how concepts are represented in the slides (e.g., through text or figures). These factors (the complexity of

the information, and the modality of information in the slides) vary locally, at shorter time scales than the one we used. Previous research has shown that factors pertaining to content can influence the coordination of multimodal behavior; for example, the use of images during classroom instruction is accompanied by gestures that help disambiguate them and aid integration between what is being said and what is being shown (Pozzer-Ardenghi & Roth, 2007). The interdependence of gesture, language, and external objects might be of such importance (Clark, 2005; Hutchins & Palen, 1997) that the best predictor of the state of each modality at any given time is the state of the other modalities, given the constraints of the specific message being conveyed and the objects available in the environment. Future work could delve deeper into behaviors at a smaller temporal grain size than the one we have chosen here, as it might be a more appropriate time scale to observe this behavior of the system.

Regarding our second objective—of examining whether modalities vary in a coordinated way, revealing system-like properties—results from both PCA solutions offer only suggestive glimpses of system-like covariation among modalities. These behavioral signals are, at least to a small extent, "moving together," suggesting that the cognitive system is partly controlling them as a coordinated unit. From a dynamic systems perspective, this would help understand the ease of multimodal coordination, as the degrees of freedom the cognitive system needs to control during the task are reduced (Kelso, 2009; Riley et al., 2011; Shockley et al., 2009). The first PCA revealed patterns in the general strategies that different speakers adopt when presenting complex information: It suggested that people who talk faster also tend to move more and talk more loudly. This is in line with some of the findings in the co-speech gesture literature indicating that gesturing facilitates fluency in speech, in particular when it involves spatial language (Morsella & Krauss, 2004; Rauscher, Krauss, & Chen, 1996). Similarly, people who have more slide changes in their presentations tend to move more. This makes sense given evidence that shows that when speakers change topics (Cassell et al., 2001) or integrate new ideas with prior discourse (Alibali et al., 2014; Enfield, 2009), they produce increased movement and gestures.

This pattern from the first PCA, in conjunction with the trends from the second PCA, has implications for the extended mind hypothesis (Clark & Chalmers, 1998). It suggests that lecture slides are integrated with the speaker's behavior and may be functioning as a part of the cognitive system itself. Studying the specific roles of the slides in the communicative process in more depth might help further clarify this possibility. Note that although multiple combinations of the modalities are possible, speakers tend to adopt and maintain a stable strategy across their whole presentation.[7] This indicates that the system is organizing itself in a way that achieves stability at longer time scales. Speaker behavior enters a particular "region of activity space" (i.e., a specific subset of all the possible combinations of the modalities), and it remains stable in that region. This is consistent with the idea that individual differences are prominent, which was also suggested by our regression analyses. Although in the present work we cannot identify the causal locus of the stabilization of behavior, a variety of factors likely contribute, such as physical constraints of the environment, and pragmatic constraints of the topic or the discourse context (including audience feedback) contribute to that process.

The second PCA analysis also reveals some slight covariation across the different modalities as they change over time. Though our results are only suggestive, if speakers subtly adapt their modalities relative to one another during a talk, this would be characteristic of soft-assembled, context-dependent systems (Kelso, 2009; Richardson et al., 2014). More broadly, and perhaps more boldly, such a synergy would support the notion that different modalities weave into each other and act as a functional group to give rise to meaning (Hutchins & Palen, 1997; Pozzer-Ardenghi & Roth, 2007). In general, the patterns from the second PCA are difficult to interpret, and some of its emerging patterns are puzzling. For instance, literature examining prosodic prominence (Cole, Mo, & Hasegawa-Johnson, 2010) shows that speech rate decreases as F0 and intensity increase. This is in contrast with some of the patterns of the second PCA (i.e., positive covariation between all three), which suggest that there might be more than prosodic prominence to the speakers' multimodal performance. One possibility is that the observed relationships are purely physical and respond to the physiology of the vocal tract: A correlational study that asked people to change the vocal effort of their speech found that when people speak softer, their fundamental frequency, and speaking rate decrease as well (Black, 1961). An alternative possibility is that the covariation reflects emotional activation during public speaking. Though still debated, studies on the acoustic properties of emotion have found that high emotional activation is usually accompanied by speech with higher fundamental frequency, higher intensity, and higher rate (see Scherer, 2003, for a review). Still another possibility is that the acoustic correlates of prosodic prominence that have been reported at the word and sentential level do not hold when looking at the dynamics of longer speech excerpts: The temporal scale of the observations is too coarse to capture more fine-grained relationships between modalities.

It is possible that additional measures related to the informational flow of the talk or the quality of the performance are needed to clarify the patterns in the data. For instance, it might be the case that effective and ineffective presenters have completely different strategies of multimodal performance and that the mixture of ability levels in the sample is making it difficult to discern the specific patterns of covariation in the speakers' multimodal behavior. Previous research has shown that individual differences in verbal working memory (Gillespie, James, Federmeier, & Watson, 2014), spatial working memory (Chu, Meyer, Foulkes, & Kita, 2014), and phonemic fluency (Hostetter & Alibali, 2007), among others, affect the type and rate of co-speech gestures speakers make. Other work has suggested that the effectiveness of a speaker's message is associated with multimodal behavioral signatures. For instance, the proportion of grooming movements and gestures that a speaker makes modulates how effective gestures are in aiding message comprehension (Obermeier, Kelly, & Gunter, 2015). For future studies, it might be interesting to obtain performance measures like learning outcomes of the audience or overall enjoyment of the presentation to test if these are associated with distinctive combination patterns of the modalities. It may be the case, for example, that speakers that show more consistency in the patterns of multimodal behavior they use achieve higher comprehension of their message.

Finally, some general caution is required when interpreting the results presented here. First, while natural datasets allow us to study behavior as it happens in the real world

and afford more ecologically valid inferences, they do so at the price of experimental control. Aspects of the video recording that were out of our control such as distance from the camera, general light level in the room, calibration of the microphone volume, and the camera brightness at each talk could have affected the results presented here. We made a deliberate effort to reduce the sources of noise by carefully selecting the videos and preprocessing the data, but it is important to acknowledge this issue. Second, while automated methods constitute great alternatives to analyze bigger datasets, they have some limitations in their accuracy. The methods used in this paper showed good levels of accuracy when compared to the performance of human coders. However, they may have introduced some noise in the signals that could have made it more difficult to detect the actual coordination of modalities. Third, we did not consider the content of these presentations, as it was outside the scope of the present analysis. Still, as we have acknowledged, the ebb and flow of slide changes, body, and voice may have been time-locked to certain aspects of the information presented in the slides or talk. Future research could use a co-registration of transcribed presentations, slides, and dynamics derived from automated methods in order to test this, although such corpora are still not widely available. Lastly, our dataset comprised talks given by mostly male experienced presenters, based in U.S. institutions, and under minimal time constraints. The trends and patterns found in this study might (and most likely would) change when dealing with other levels of experience in presentations, more gender-balanced samples, more geographically and culturally diverse samples, and different time constraints. Exploring these variables and their effects on multimodal signal control is an interesting direction for future research.

5. Conclusion

The use of different modalities of communication during complex information presentation exhibits some weak but interesting regularities. System-like compression is observed in the way the different modalities are coordinated. Body, speech, and slides may begin to approximate an integrated unit during communicative performance across the presentation, giving some support to the idea of an extended cognitive system. Further exploring individual differences and including measures of informational complexity could help further clarify the specific coordination synergies that speakers engage in during multimodal communication.

Acknowledgments

This project was supported in part by the Graduate Dean's Recruitment Fellowship awarded by UC Merced to the first author. It has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement no. 705037 to A.G., who is now at the Department of Psychological Science at the University of North Carolina at Charlotte. We thank Dr.

Anne Warlaumont and Dr. Chris Kello for their helpful comments during the development of this project and their suggestions on earlier versions of this manuscript.

Notes

1. www.youtube.com/channel/UCRcuWjRqxZ2RHvEdZGAliWw.
2. https://vimeo.com/user8418321.
3. The R scripts and datasets we used in this project, as well as additional figures, are publicly available in GitHub: https://github.com/camialviar/ComplexCommunicationDynamics.
4. This last component could reflect sex differences in pitch. This was not a variable of interest to us, and so we do not test this in our dataset. However, it is also important to note that some males exhibited perceptually higher pitches, so the relationship suggested by component 3 between pitch and intensity might not be exclusively due to sex.
5. We set alpha for p to 0.05. By chance, one would anticipate no more than about one of our coefficients to yield significance. Four significant coefficients suggest our results are not due to type I error, though it is important to note that the observed effects are small and future analyses, especially where models might become more complex, would benefit from a correction for alpha.
6. We examined whether this peak reflects an increase in graph usage, following a reviewer's suggestion. We calculated the percentage of slide changes involving graphs for five speakers, for whom window 8 had a big peak. The results suggest that the peak was not driven by an increase in the percentage of graphs being used between window 7 (M = 47.44%) and window 8 (M = 49.31%). A detailed content analysis would be required to better clarify the variables driving the increase.
7. For the interested reader, figures showcasing the clustering of the data when projected in the first two components of the first PCA solution are available in the GitHub repository for the project: https://github.com/camialviar/ComplexCommunicationDynamics.

References

Abrahams, M. (2016). A big data approach to public speaking. *Insights*. Available at https://www.gsb.stanford.edu/insights/big-data-approach-public-speaking. Accessed February 14, 2019.

Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and Instruction*, *32*(1), 65–100. https://doi.org/10.1080/07370008.2013.858161.

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.), *AVSP* (pp. 173–177). Queensland: ISCA Archive.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). Cicero-towards a multimodal virtual audience platform for public speaking training. In: R. Aylett, B. Krenn, C. Pelachaud & H. Shimodaira (Eds.), *International workshop on intelligent virtual agents* (pp. 116–128). Edinburgh: Springer. https://doi.org/10.1007/978-3-642-40415-3_10

Black, J. W. (1961). Relationships among fundamental frequency, vocal sound pressure, and rate of speaking. *Language and Speech*, *4*(4), 196–199.

Boersma, P., & Weenink, D. (2010). Praat: Doing phonetics by computer (Version 5.1.31) [Software]. Available at http://www.praat.org. Accessed February 14, 2019.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7–9), 1044–1098.

Brock, S., & Joglekar, Y. (2011). Empowering PowerPoint: Slides and teaching effectiveness. *Interdisciplinary Journal of Information, Knowledge, and Management*, *6*(1), 85–94.

Cassell, J., & McNeill, D. (1991). Gesture and the poetics of prose. *Poetics Today*, *12*(3), 375–404. https://doi.org/10.2307/1772644.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *7*(1), 1–34. https://doi.org/10.1075/pc.7.1.03cas.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). Non-verbal cues for discourse structure. In B. L. Webber (Ed.), *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 114–123). Toulouse: Association for Computational Linguistics.

Chen, L., Leong, C. W., Feng, G., & Lee, C. M. (2014). Using multimodal cues to analyze mla'14 oral presentation quality corpus: Presentation delivery and slides quality. In X. Ochoa, M. Worsley, K. Chiluiza, & S. Luz (Eds.), *Proceedings of the 2014 ACM workshop on multimodal learning analytics workshop and grand challenge* (pp. 45–52). Istanbul: ACM.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694. https://doi.org/10.1037/a0033861.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243–268). Hillsdale, NJ: Erlbaum.

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, *7*(4–5), 507–525. https://doi.org/10.1177/1461445605054404.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*, 7–19. https://doi.org/10.1093/analys/58.1.7.

Cohen Priva, U. C. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, *160*, 27–34. https://doi.org/10.1016/j.cognition.2016.12.002.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*(2), 425–452. https://doi.org/10.1515/labphon.2010.022.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515–535.

Enfield, N. J. (2009). *The anatomy of meaning: Speech, gesture, and composite utterances* (Vol. 8). Cambridge, UK: Cambridge University Press.

Enfield, N. J. (2013). *Relationship thinking: Agency, enchrony, and human sociality*. Oxford, UK: Oxford University Press.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51. https://doi.org/10.1016/j.jml.2009.09.002.

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451. https://doi.org/10.1080/01690965.2013.796397.

Gamer, M., Lemon, J., & Fellows, I. (2012). irr: Various coefficients of interrater reliability and agreement. R package version 0.84. Available at https://CRAN.R-project.org/package=irr. Accessed February 14, 2019.

Gillespie, M., James, A. N., Federmeier, K. D., & Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, *132*(2), 174–180. https://doi.org/10.1016/j.cognition.2014.03.012.

Goodwin, C. (2000). Action and embodiment within human situated interaction. *Journal of Pragmatics*, *32*, 1489–1522. https://doi.org/10.1016/S0378-2166(99)00096-X.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Second international conference on spoken language processing* (pp. 429–432). Alberta: ISCA Archive.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664.

Hellermann, J. (2003). The interactive work of prosody in the IRF exchange: Teacher repetition in feedback moves. *Language in Society*, *32*(1), 79–104. https://doi.org/10.10170S0047404503321049.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*(4), 575–591. https://doi.org/10.1016/j.system.2005.04.002.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. In M. Home (Ed.), *Prosody: Theory and experiment* (pp. 335–350). Dordrecht: Springer. https://doi.org/10.1007/978-94-015-9413-4_12

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In A. W., Biermann (Ed.), *Proceedings of the 24th annual meeting on Association for Computational Linguistics* (pp. 136–144). New York: Association for Computational Linguistics.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79*, 1–17. https://doi.org/10.1016/j.jml.2014.10.004.

Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, *7*(1), 73–95. https://doi.org/10.1075/gest.7.1.05hos.

Hutchins, E. (1983). Understanding Micronesian navigation. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 191–225). Hillside, NJ: Erlbaum.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*(3), 265–288. https://doi.org/10.1207/s15516709cog1903_1.

Hutchins, E., & Palen, L. (1997). Constructing meaning from space, gesture, and speech. In L. B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, tools and reasoning* (pp. 23–40). Berlin: Springer.

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In A. Sarkar, C. Rose, S. Stoyanchev, U. Germann, & C. Shah (Eds.), *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 638–646). Boulder, CO: Association for Computational Linguistics.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223–232. https://doi.org/10.1016/j.tics.2010.02.005.

Kelso, J. S. (2009). Synergies: Atoms of brain and behavior. In D. Sternad (Ed.), *Progress in motor control* (pp. 83–91). Boston: Springer.

Kirsh, D. (2010). Thinking with external representations. *AI & Society*, *25*(4), 441–454.

Koppensteiner, M., & Grammer, K. (2010). Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, *44*(3), 374–379. https://doi.org/10.1016/j.jrp.2010.04.002.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Levasseur, D. G., & Kanan Sawyer, J. (2006). Pedagogy meets PowerPoint: A research review of the effects of computer-generated slides in the classroom. *The Review of Communication*, *6*(1–2), 101–123. https://doi.org/10.1080/15358590600763383.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45–66.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356. https://doi.org/10.1016/j.tics.2010.06.002.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. (2000). *Language and gesture* (Vol. 2). Cambridge, UK: Cambridge University Press. Available at http://books.google.com/books?hl=en&lr=&id=DRBcMQuSrf8C&oi=fnd&pg=PR9&dq=gesture+mcneill&ots=jBDQ3Cwqqt&sig=mGXmcSQeJbs2-goD-1SeSF0CpVU. Accessed February 14, 2019.

McNeill, D. (2008). *Gesture and thought*. Chicago: University of Chicago Press.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, *4*(2), 119–141. https://doi.org/10.1075/gest.4.2.02mel.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.

Morsella, E., & Krauss, R. (2004). The role of gestures in spatial working memory and speech. *The American Journal of Psychology*, *117*(3), 411–424. https://doi.org/10.2307/4149008.

Moulton, S. T., Türkay, S., & Kosslyn, S. M. (2017). Does a presentation's medium affect its message? PowerPoint, Prezi, and oral presentations. *PLoS ONE*, *12*(7), e0178774. https://doi.org/10.1371/journal.pone.0178774.

Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social Cognitive and Affective Neuroscience*, *10*(9), 1236–1243. https://doi.org/10.1093/scan/nsv011.

Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, *45*(2), 329–343. https://doi.org/10.3758/s13428-012-0249-2.

Pouw, W., & Dixon, J. A. (2018). Quantifying gesture-speech synchrony: Exploratory data report and preregistration. *Open Science Framework*. https://doi.org/10.31234/osf.io/983b5

Pouw, W., Trujillo, J., & Dixon, J. A. (2018). The quantification of gesture-speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Open Science Framework*. https://doi.org/10.31234/osf.io/jm3hk

Pozzer-Ardenghi, L., & Roth, W. M. (2004). Photographs in lectures: Gestures as meaning-making resources. *Linguistics and Education*, *15*(3), 275–293. https://doi.org/10.1016/j.linged.2005.01.001.

Pozzer-Ardenghi, L., & Roth, W. M. (2007). On performing concepts during science lectures. *Science Education*, *91*(1), 96–114. https://doi.org/10.1002/sce.20172.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rauscher, F., Krauss, R., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, *7*(4), 226–231. https://doi.org/10.1111/j.1467-9280.1996.tb00364.x.

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). Complex dynamical systems in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253–282). New York: Cambridge University Press.

Riley, M. A., Richardson, M., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in Psychology*, *2*, 38. https://doi.org/10.3389/fpsyg.2011.00038.

Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. *Ninth European conference on speech communication and technology* (pp. 513–516). Lisboa: ISCA Archive.

Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 308–329). Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781118325001.ch15

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169–182.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5.

Scherer, S., Layher, G., Kane, J., Neumann, H., & Campbell, N. (2012). An audiovisual political speech analysis incorporating eye-tracking and perception data. In N. Calzolari, et al. (Eds.), *LREC* (pp. 1114–1120). Istanbul: European Language Resources Association.

Schmidt, C. P., Andrews, M. L., & McCutcheon, J. W. (1998). An acoustical and perceptual analysis of the vocal behavior of classroom teachers. *Journal of Voice*, *12*(4), 434–443. https://doi.org/10.1016/S0892-1997(98)80052-0.

Selting, M. (2010). Affectivity in conversational storytelling. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *20*(2), 229–277.

Shim, H. S., Park, S., Chatterjee, M., Scherer, S., Sagae, K., & Morency, L.-P. (2015). Acoustic and paraverbal indicators of persuasiveness in social multimedia. In V. Clarkson, & J. Manton (Eds.), *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2239–2243). Queensland: IEEE.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319. https://doi.org/10.1111/j.1756-8765.2009.01021.x.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3.

Tseng, C. Y., Pin, S. H., Lee, Y., Wang, H. M., & Chen, Y. C. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, *46*(3), 284–309. https://doi.org/10.1016/j.specom.2005.03.015.

Valbonesi, L., Ansari, R., McNeill, D., Quek, F., Duncan, S., McCullough, K. E., & Bryll, R. (2002). Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures. In EUSIPCO (Ed.), *11th European signal processing conference*. Toulouse: IEEE.

Van der Maaten, L. J. P., Postma, E. O., & Van Den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009. Available at https://lvdmaaten.github.io/drtoolbox/. Accessed February 14, 2019.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B*, *369* (1651), 20130292. https://doi.org/10.1098/rstb.2013.0292.

Voigt, R., Podesva, R. J., & Jurafsky, D. (2014). Speaker movement correlates with prosodic indicators of engagement. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Speech prosody* (Vol. 7). Dublin: ISCA.

Vyas, P., & Sharma, S. (2014). A study on the efficacy of PowerPoint for writing instruction. *International Journal of Instructional Technology & Distance Learning*, *11*(8), 29–42.

Wehling, E. (2018). Discourse management gestures. *Gesture*, *16*(2), 245–276. https://doi.org/10.1075/gest.16.2.04weh.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*(10), 1559–1579. https://doi.org/10.1016/j.pragma.2005.04.012.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968.

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings of the international conference: From sound to sense* (pp. 97–102). Cambridge, MA: MIT.