The behavior of hand and facial gestures at pauses in speech

Julia Myers

Yale University
Jelena Krivokapic, advisor
May 1, 2012

Table of Contents

Abstract

Previous studies have presented conflicting findings as to the nature of gesturing at pauses, and as a result have drawn a wide variety of conclusions about the function of gesture in relation to speech. These conflicting findings have primarily been due to studies' differing conceptions of what constitutes a "gesture" as well as what part of the gesture movement is significant and worth measuring. For example, some studies examine only hand gestures while others examine only facial gesture; some studies examine only the start of a gesture, while others examine only the end of a gesture. This study aims to resolve discrepancies about gesture behavior at pauses by conducting a more inclusive gesture study observing both hand and facial gestures, as well as marking the onset (start), target (end), and duration of each gesture. This study also seeks to posit potential gestural indications of grammatical versus ungrammatical pauses. Finally, this study seeks to improve accuracy and increase efficiency of gesture labeling for future studies.

The present study elicits gesturing from six subjects via spontaneous, monologue speech. Gesturing was labeled in pauses, in the region immediately surrounding pauses, and in a corresponding amount of fluent speech. The results of this study indicate that in pauses, full gestures do not occur often, but parts of previous gesture or upcoming gesture often bridge into or out of the pause, most notably out of the pause. Gesture suspensions, indicated by a hold or "freezing" effect, are prevalent in pause areas, often starting in the pre-pause region and ending in the pause. Gesture behavior in grammatical versus ungrammatical pauses did not indicate a distinctive pattern. Lastly, an improved method for labeling gestures was developed using motion tracking of video data in combination with software that automatically detects gesture landmarks based on velocity information.

1. Introduction

Studies that have examined bodily gesture reach a wide variety of conclusions as to the nature of gesture in relation to speech. Some hypothesize that the two occur in synchrony and are part of an integrated cognitive system (see Sassenberg et al 2010), while others hypothesize that gesture is a paralinguistic phenomenon serving to aid speech production, and occurs more often when speech becomes difficult (see Moscovici 1967, Werner and Kaplan 1963). These theories however are based on results from studies that examine different parts or types of gesture behavior. This study therefore aims to (1) improve upon past studies by observing all components of gesture behavior, and (2) contribute to the dialogue about gesture behavior by focusing on the behavior of gesture in the absence of speech, i.e. at pauses, with the goal of better understanding gesture behavior in relation to speech.

2. Background

2.1 Fluent and disfluent pauses

Pauses have often been divided into two categories, fluent and disfluent (also, grammatical and ungrammatical, respectively). Fluent pauses refer to cessations of speech that occur at prosodic boundaries. These pauses are perceived as intended breaks between meaningful chunks of speech, and have often been regarded as locations of speech planning (Cooper and Paccia-Cooper 1980, Ferreira 1991). Conversely, disfluent pauses refer to cessations of speech that can occur anywhere within the utterance, including at normally predicted boundaries, and result from a breakdown in the relation of speech content. These pauses are not considered to be planned pauses.

Currently there are no reliable acoustic or articulatory indicators allowing one to distinguish between a fluent and disfluent pause. As such, this study will include all pauses regardless of perceived fluency or disfluency. Perceived pause fluency or disfluency will be noted at a later point in the study.

Historically the technical term “pause” also includes a category known as filled pauses. These refer to any pause that contains a filler word; for instance, “um,” and “er.” This study will examine all silent pauses in speech including the silences surrounding filler words; it will not however include the filler word as part of the pause, because

fillers do not involve actual cessation of the speech articulators and as such are better investigated separately from silent pauses. Because filled pauses are typically perceived as disfluent pauses, those silent pauses containing a filler word will be appropriately filtered in the pause type perception portion of the study. Future studies should however also examine filled pauses, including both the filler word and the silent pauses surrounding the filler, as a category completely separate from silent pauses.

2.2 Nonarticulatory gesture

2.2.1 Gesture types

"Gesture" is an umbrella term for a wide variety of body movements. In speech production studies, the term refers to articulatory gestures, which are linguistically relevant movements of the speech apparatus such as tongue tip movement and lip opening and closure. Outside speech production studies, the term "gesture" commonly refers to a broader set of nonarticulatory movements made by the face, hands, and body. One of the pioneers in gesture research, Adam Kendon, categorized body gestures according to a spectrum, which would later be coined "Kendon's continuum" by a fellow researcher, David McNeill (see Kendon 2004:104):

Gesticulations → Language-like Gestures → Pantomimes → Emblems → Sign Languages

This continuum consists of both a categorization and an ordering of nonverbal gesture. As one moves from left to right on the spectrum, gestures become less reliant on accompanying speech, their language-like features increase, and idiosyncratic gestures become replaced by socially regulated forms (McNeill 1992: 37).

The leftmost category, gesticulation, consists of idiosyncratic hand and facial movements that occur spontaneously with accompanying speech. Language-like gestures are nearly identical to gesticulations but are crucially connected to the grammatical and semantic meaning of the corresponding sentence; for instance, a language-like gesture could be used in lieu of a spoken adjective: "It was quite [gesture]," with the arms thrown back and the eyebrows raised to indicate "scary" or "terrifying." Pantomime gestures involve the portrayal of certain behaviors, objects, or ideas with the body that do not

require accompanying speech, and can be used in sequence to create meaning. Emblems are gestures that have been codified by society and possess a highly regular form. Examples of emblems in American English include putting the thumb and pointer finger together to indicate the sign for "OK," or turning up the thumbs to indicate a job well done. These can be understood without accompanying speech, and are much more constrained in form than pantomime. Last, sign languages are complex systems of hand and facial gesture that possess all the structural properties of spoken language. Those who sign also use separate language accompanying hand and facial gestures.

Gesticulations have been the focus of many gesture/speech studies, as they are the least influenced by external regulation of form and occur in the context of accompanying speech, as opposed to serving as a replacement for language. Gesticulations have been divided into further categories, most commonly iconic, metaphoric, deictic, and beat gestures (McNeill 1992). According to this classification, iconic gestures depict concrete entities (such as a "bowl" with hands cupped in supine position), metaphoric gestures depict abstract concepts as if they had form (such as "liberty" with the arms flying out), deictic gestures indicate location and orientation (such as pointing to an actual object or to a represented concept in space), and beat or "baton" gestures are rhythmic ups and downs made by the hand and face that time with speech and may indicate emphasis of particular information. For the purposes of this study, the term "gesture" will refer to the aforementioned gesticulations.

2.2.2 Gesture onset and target

All movements have a start and end point; this includes gesture movements. The start point, referred to as the gesture onset, is the moment the head or hand leaves rest position or embarks in a new direction from a previous movement (McNeill 2000: 204). The end point, referred to as the gesture target, is the moment a gesture reaches its destination, as indicated by a cessation of motion or a changed direction (Ibid.). The time between the onset and target of a gesture is the gesture duration.

In the present study, gesture onset, target, and duration are noted for each distinct movement observed in the hand or face. This allows for separate observation of gesture start and gesture end, as well as complete gesture units (the interval from start to end).

2.2.3 Gesture phases

McNeill and Levy (1982) proposed that gestures can be broken into three phases: a preparatory phase, the stroke phase (the "actual gesture"), and a retraction or release phase (see Figure 1).

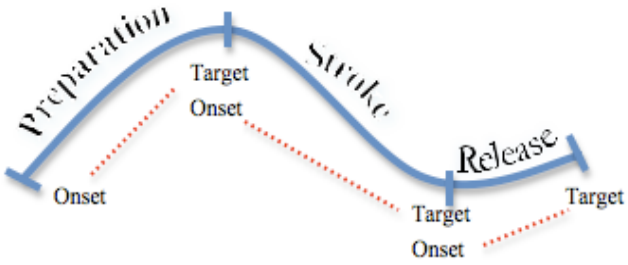


Figure 1. The three gesture phases

The stroke has been taken to represent the primary intent of the gesture, whether a discrete hit of the hands or a continuous representation of an entity such as a specific shape. The preparation therefore serves only to bring the head or hands to the appropriate position to execute the gesture stroke. The release has been taken as a natural return to rest position following the stroke phase.

Typically, the target of a preparation phase is identical to the onset of a stroke phase, and similarly, the target of a stroke phase is identical to the onset of a release phase; however, if the speaker interrupts his gesture with a temporary hold, there may be time between the phases, or an interruption in the middle of a phase. Not all three phases may be completed with every gesture, primarily in cases where gestures flow smoothly from one motion into another.

Labeling gesture phases, i.e. grouping movements into larger gestures based on the apparent relationship between each movement, is an inherently subjective process. As such, the three gesture phases proposed by McNeill and Levy will be noted in the following study, but separately from objective onset and target labels for each distinct movement.

2.2.4 Gesture suspensions

Sometimes, a gesture is interrupted before reaching its apparent target. These interruptions, termed gesture suspensions (by researcher Mandana Seyfeddinipur at the Max Planck Institute for Psycholinguistics in Nigmegen, The Netherlands), can be in the form of holds and freezes, where the gesture hovers or freezes above rest position in the manner of the previous gesture configuration, or premature releases, where the gesture falls back to rest position before reaching its target. Gesture suspensions thus refer both to the lack of gesture and the fact that the lack of gesture occurs between surrounding gesture movements (see Seyfeddinipur 2006 for more discussion).

2.3 Controversy: nonverbal gesture in pauses

The behavior of gesture in pauses is not well understood. This is primarily due to the wide range of behaviors "gesture" can denote, the wide range of categories "pause" can denote, and the resulting discrepancies in studies examining such behaviors. For example, in 1978 Butterworth and Beattie found that "gestures" occur more often in pauses than in fluent speech; however, their "gestures" referred to gesture onset, with no discussion as to gesture target location. Conversely, other studies have noted gesture behavior primarily in fluent speech, finding for instance that beat gesture targets or "hits" appear to align closely with pitch accents (Tuite 1993, McClave 1997, Cave, Guaitella, Bertrand, Santi, Harlay and Espesser 1996, Keating 2003, Yasinnik, Renwick, and Shattuck-Hufnagel 2004, Loehr 2004) and affect perception of prominence (Treffner, Peter, and Kleidon 2008, Krahmer and Swerts 2007); however these studies examined gesture target exclusively, making no mention of corresponding onset location or behavior (see also Cassell, McCullough, and McNeill 1999).

These differing observations based on different components of gesture, among other discrepancies between studies, have caused divergent theories about gesture function in relation to speech. One theory, the "difficulties" view (see Sassenberg et al. 2010), is based on the result that gestures occur more often in pauses and during disfluent breaks in speech. This theory posits that gestures increase with increased speech task difficulty (Sassenberg et al. 2010), and serve to aid the speaker in lexical search or in the planning of speech when speech stops or falters, i.e. at pauses (see Moscovici 1967,

Werner and Kaplan 1963). Another theory, the "gestures-as-simulated-action" (GSA) view, is based on the result that gestures occur more often in fluent speech and less often in the absence of speech, i.e. at pauses. This theory proposes that gestures are a by-product of mental imagery, and increase with increased mental imagery (Sassenberg et al. 2010). Another theory, the "integrated system" framework (Mayberry and Jacques 2000 in McNeill 2000), is based on the findings that parts of gesture temporally align with certain speech properties, and posits that the gesture and speech systems are jointly planned and integrated before execution (Kendon 1980, McNeill 1985, 1992).

2.4 Research question and hypothesis

This study aims to determine gesture behavior at pauses in speech compared with gesture behavior in fluent speech. Pilot studies conducted in preparation for the current study indicate that gesturing does not occur often in pause centers, although gesturing does occur at pause edges, most commonly bridging out of the pause. In these cases, gesture onset occurs in the pause and gesture target occurs in following speech, often aligning with an intonation peak (i.e. pitch accent, see section 2.3 for further discussion). Therefore, the hypothesis for the present study is that gestures will not occur often in pauses, but may occur bridging into or out of the pause; additionally, in or around silent pauses, gesture suspensions will be more prevalent than in fluent speech. Finally, regarding disfluent and fluent pauses, the hypothesis is that gesture suspensions will be more common in pauses perceived as disfluent, and that gestures bridging from previous speech will be more common in pauses perceived as fluent. This is due to the abrupt and unplanned nature of the gesture suspension, which parallels the abrupt and unplanned nature of the disfluent pause, and the fluid nature of the gesture completion in the pause, which parallels the fluid nature of the fluent pause.

3. Experiment

3.1 Introduction

This study seeks to improve upon past studies by including all gesticulation types, observing both hand and facial gesturing, and marking every gesture with its onset, target, duration, description of behavior, and apparent gesture phase. This study examines all silent pauses including silent areas surrounding filler words, and labels perceived pause fluency and disfluency separately. It examines monologue speech in order to eliminate labeling difficulty associated with dialogue speech overlaps.

3.2 Methods

3.2.1 Subjects

Seven subjects, five female (F1-F5) and two male (M1 and M2), participated in the experiment. Subject F1 was excluded due to poor audio, with six subjects remaining. All subjects were native speakers of American English, naïve as to the purpose of the experiment, and paid for their participation.

3.2.2 Data collection

Subject F2 was filmed on a Sony HD150 video camera. M1 was filmed on a FlipVid Ultra HD video camera. Subjects F3, F4, F5, and M2 were filmed on a Canon 7D. Subjects F2 and F3 were recorded at 24 frames per second. Due to the development of a new potential labeling system involving motion-tracking data (discussed in Section 6), the remaining subjects were recorded at 60 frames per second to enhance motion tracking capability. Audio was recorded separately on a Marantz audio recorder with an attached shotgun microphone. The microphone was placed on a stand and pointed at the subject's mouth at about a foot-and-a-half distance. The video camera was set on a tripod directly in front of the subject, at about a six-foot distance, and pointed at the front of the subject's body to capture visuals of speech and gesturing. The experimenter stood directly to the left of the camera. The subject was told to speak to the experimenter, not the camera. The experimenter listened to the subject but did not actively participate in a dialogue discussion with the subject.

3.2.3 Stimuli

The question chosen for the present study was, "Do you think the elephant or the zebra will go extinct first; and why?" Subjects were then asked to debate the question out loud until they settled on an answer. There was no time limit given for responses.

The spontaneous question chosen for the present study was excerpted from a larger experiment examining primarily semi-spontaneous, semi-controlled speech (not discussed in this paper). Four subjects, F2, F3, F5, and M1, participated in the larger experiment while the remaining two subjects, F4 and M2, participated only in answering the single spontaneous question. The question was written on a single PowerPoint slide and displayed on a 16'' Macbook Pro laptop positioned a few feet from the subject, and out of camera view.

3.2.4 Fluent vs. disfluent pause perception study

In order to compare gesture behavior in pauses perceived as fluent or disfluent, one naïve listener, a native speaker of American English, was played the speech excerpts from all six subjects above. For a given subject, the entire speech excerpt was first played to give the listener a sense of speech rate as well as content. Then, areas of speech containing a marked pause were played through one at a time. Enough surrounding speech was included with each pause to give the listener a sense of the pause's placement in speech. The listener was then asked to categorize each pause as unnatural or natural sounding. The listener's responses were noted next to each pause.

The listener's responses closely matched the experimenter's own pass through, indicating consistency in perception of pauses as fluent or disfluent. Only the listener's responses were used for the purposes of data analysis.

3.2.5 Data analysis

Audio and video files were imported separately into Final Cut Pro, a nonlinear video editing program, and then synchronized together so that each audio clip matched with each video. The video portions of the synchronized files were exported in H.264 high quality format. They were then imported separately by subject into ELAN, a gesture and linguistic annotation program (EUDICO Linguistic Annotator, developed at the Max

Planck Institute for Psycholinguistics, freely downloadable at http://www.lat-mpi.eu/tools/elan/). In order for ELAN to read audio files and utilize certain audio recognition plug-ins, audio files must be in .WAV format and imported separately. Consequently, audio portions of the synchronized files were exported separately from Final Cut Pro, converted from .AIFF to .WAV format in Audacity, an audio recording and editing software (developed in 1999 by Dominic Mazzoni and Roger Dannenberg at Carnegie Mellon University, freely downloadable at http://audacity.sourceforge.net/), and imported into the corresponding subjects' project files in ELAN. These .WAV audio files were later imported into Praat for speech analysis (developed by Paul Boersma and David Weenink, University of Amsterdam, freely downloadable at www.praat.org).

3.2.5.1 Speech labeling

The audio file for each subject was first analyzed in ELAN, via an Audio Recognizer plug-in, for preliminary pause detection. An Audio Recognizer found silences at a minimum duration of 200ms (most studies have specified a minimum duration of "silence" for an acceptable pause, e.g., not less than 200 ms (Rochester, 1973)). All pause regions with durations greater than 200ms were detected and labeled automatically. The Audio Recognizer demonstrated consistency and reliability. The information was then exported to Praat for refinement and addition of speech transcription.

In Praat, pauses were refined manually to filter out potential errors from the automated process, for instance, the inclusion of a very soft utterance in a pause section. The word immediately before and immediately after each pause was also labeled, and in a separate tier the word was labeled as either "Pre" or "Post" the pause region. If a disfluent segment of speech occurred immediately before or after a pause, it and the word next to it were included in the pre- or post-pause region. If only one word came between two pauses, it was labeled as "Between."

Acoustic transcriptions, including refinement of the detected pauses, were based on spectrogram analysis. Pauses were indicated by a lack of vocal tract activity in the spectrogram, i.e. no signs of formant structure or frication. Surrounding speech was indicated by the presence of formant structure or frication. Listening provided a general

guide as to where to place the boundary on a given word, and trail-offs of the lowest formants or of frication determined refined cutoff points for a given word label. An additional tier was created, "Perceived Disfluencies/Miscellaneous," for notations of filled pauses ("ums" and "ers") as well as any breaths or perceived disfluent sections.

The total duration of all pause regions for a given subject was calculated, and then approximately the same amount of fluent speech was marked off for gesture labeling. The criteria for fluent speech was that it contain no marked disfluencies, no pauses, no fillers, no pre-pause or post-pause regions, and that it sound fluid and natural. These marked off sections of fluent speech were then labeled as "Fluent Area"s. Any other regions were labeled "Unmarked."

3.2.5.2 Gesture labeling

Two subjects (F4 and M2) were labeled for hand gesturing, while four subjects (F2, F3, F5, M1) were labeled for facial gesturing. Subjects were labeled according to whichever kind of gesturing, hand or facial, was used more prominently and was thus clearer to label.

3.2.5.2.1 Obscuring speech information

As facial gestures are so close in proximity to the specific articulatory gestures involved in speech production, a method was developed to remove speech information from the face so as to allow for better, unbiased labeling. This method involved motion tracking the subject's head and applying a mask over the mouth that followed its motion (see Figures 2 and 3). In this way speech information was obscured without perception of broader head gestures affected. Motion tracking and mask application was done in Adobe After Effects, a video effects software, and the footage was exported to ELAN for gesture labeling. The two subjects labeled for hand gestures were not masked. Subject F2 was labeled before this method was developed, so no mask was added to F2's footage. In addition to this visual masking method, accompanying audio was muted in ELAN so as not to influence gesture labeling.

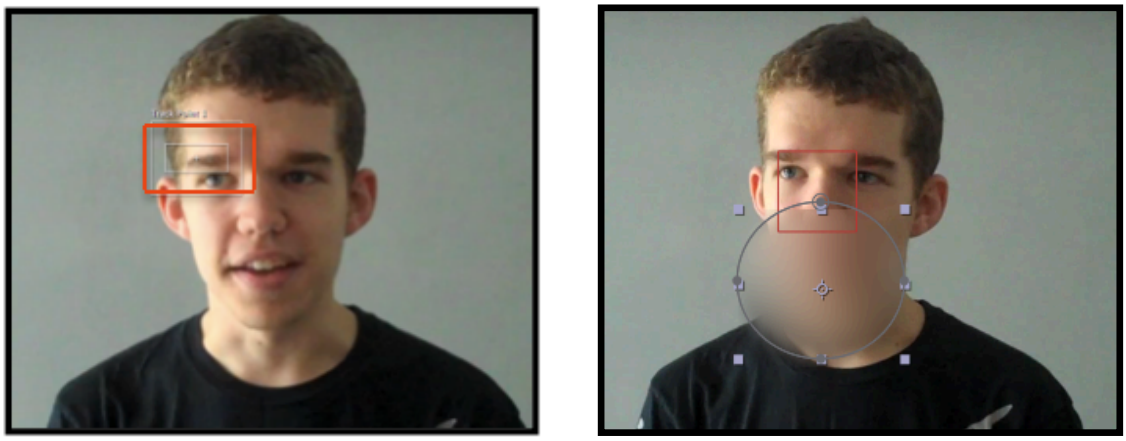


Figure 2. M1 mask application. Left image: Eyebrow position is tracked through time. Box around eye represents tracker search region, highlighted in red here. Right image: Blur mask applied to mouth. Using tracker information from a, the mask moves in synchrony with the subject's head, covering the subject's mouth and throat at all times

Figure 3. F5, final result of speech-obscuring mask

3.2.5.2.2 Gesture annotations

Gestures were labeled for onset, target, duration, and description of movement (see Figures 4 and 5). Again, onset refers to the moment the head or hand leaves rest position or embarks in a new direction from the previous movement, and target refers to the moment a gesture reaches its destination, as indicated by a cessation of motion or a changed direction (McNeill 2000: 204). Thus, if two gesture movements occur one immediately after the other, the target of the first movement and the onset of the second movement will be identical.

The video of each subject was played frame by frame. Onsets were marked according to when the head or hand started moving from rest position or started in a new direction from a previous movement. Targets were marked according to when the head or hand stopped moving and was followed by either a cessation of movement or a changed direction. In cases where a changed direction followed the movement, the moment of stopped motion was no more than a single frame; this zero velocity moment was marked as the target (based off of speech-accompaying gesture labeling methods in Yasinnik, Shattuck-Hufnagel, and Veilleux 2005). In cases where a continuous gesture occurred, such as a fluid, repetitive circling of the hands, the gesture onset was labeled as the start of the first cycling movement, and the target was labeled once the cycling stopped and the hands ceased movement or changed their motion path. The description of the gesture involved notating a brief reference as to the nature of the movement, e.g. "Head up" or "Head to left." Self-adjusting gestures were not included, i.e. scratching the neck or touching one's hair.

In a separate tier, gestures were also classified based on observation as one of the three phases proposed to be part of a larger, unified gesture: preparation, stroke, and release (McNeill and Levy 1982) (see Figures 4 and 5). For instance, "Head up" could be labeled as preparation or release depending on its apparent relation to a nearby stroke. Additionally, several gestures could be grouped together and labeled as a single stroke, if the gestures appear to express the same idea or concept.

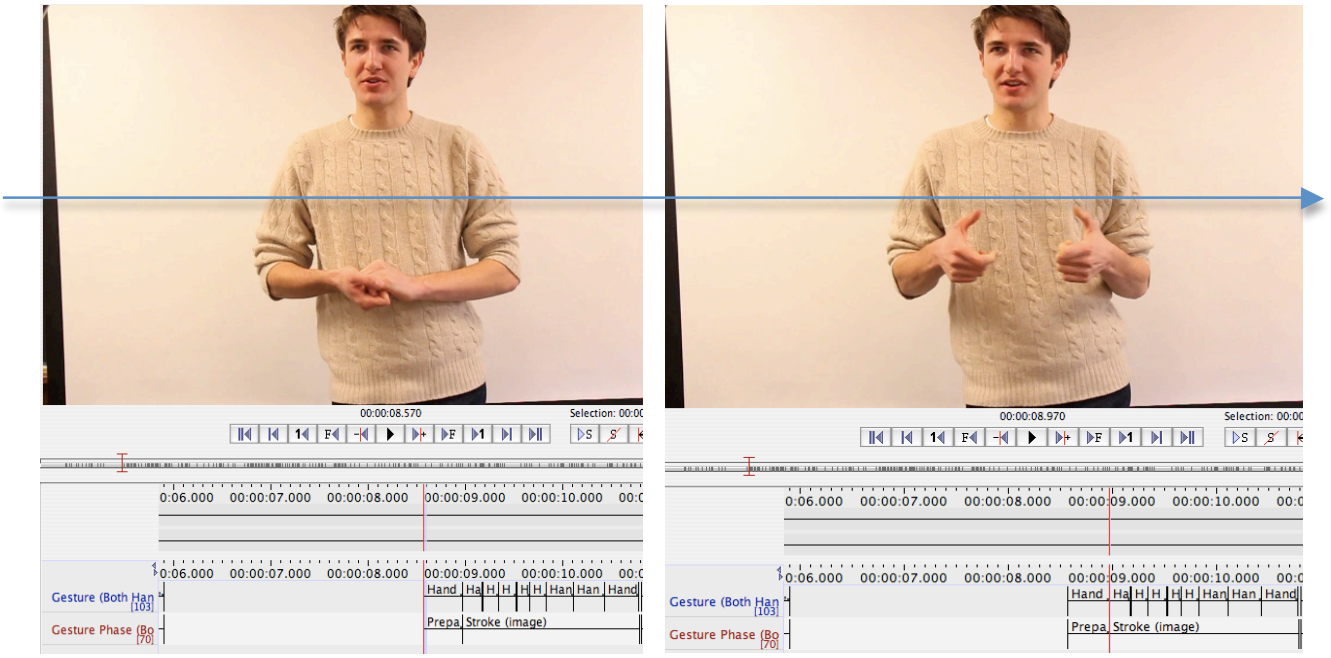


Figure 4. M2 example of gesture onset and target, separately labeled as preparation phase

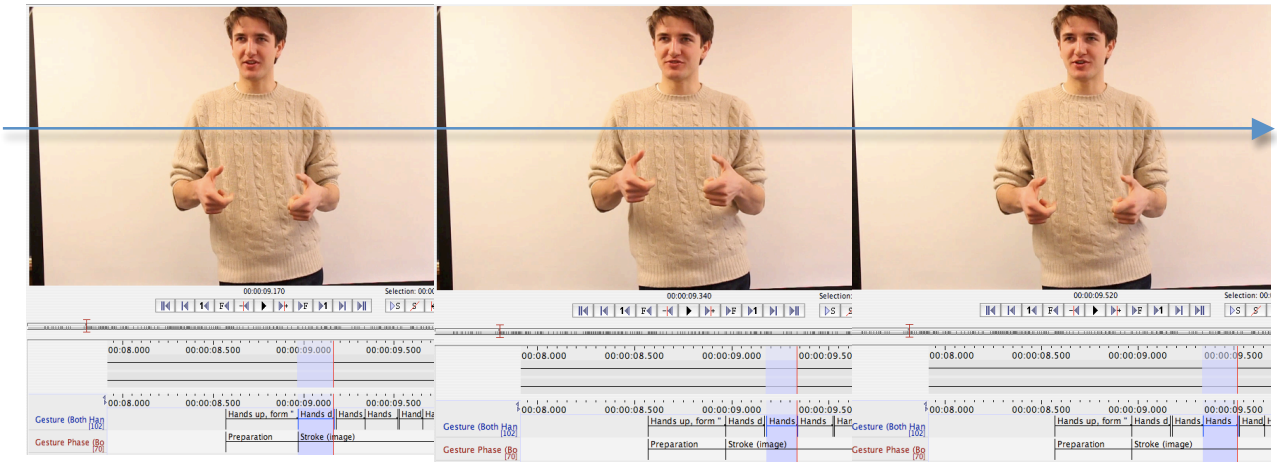


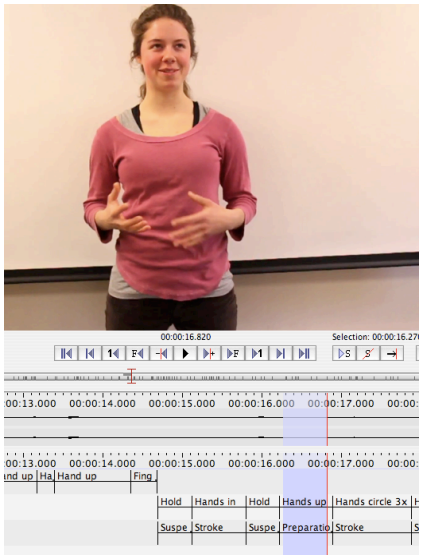
Figure 5. M2 example of multiple gestures grouped into a unified stroke phase. Note hand distance from blue arrow. Here, the subject makes repeated motions of a book shape, all motions indicating the same "main idea"

3.2.5.2.3 Gesture suspensions

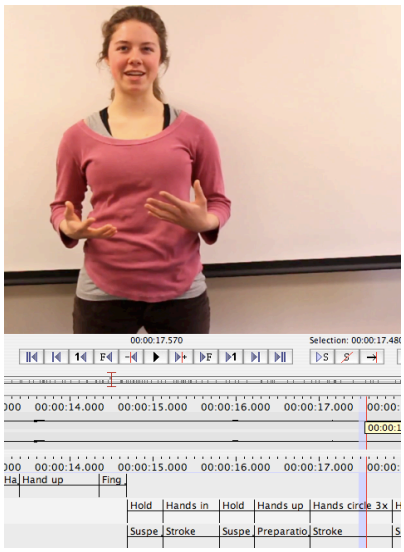
In cases where a gesture was held before reaching its apparent target, the activity was labeled as a gesture suspension in the gesture phase tier. Suspensions had three varieties (as discussed in Section 2.2.4): holds, freezes, and premature releases. Suspensions of the hold variety often involved some motion but the velocity of the motion was negligent compared to velocities of surrounding gestures. Suspensions of the

freezing variety did not involve any motion, and additionally had the appearance of an actual freeze of the body part as opposed to a hovering behavior. Suspensions of the premature release variety involved motion of the head or hands slackening or suddenly dropping in the middle of motion. Figure 6 illustrates two suspensions of the hold variety in the context of other gesture movements.

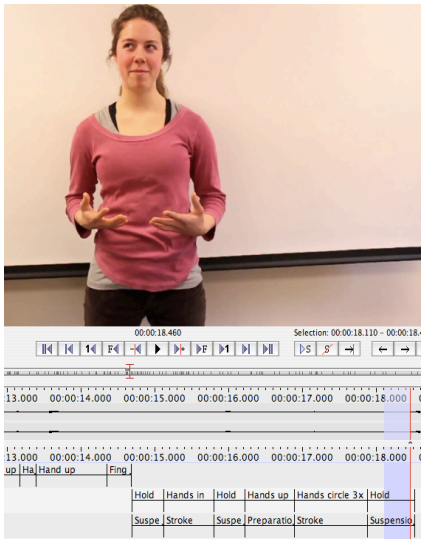
a. Hands hold (Suspension)



b. Hands start up (Preparation)



c. Hands go down (Stroke)



d. Hands hold (Suspension)

Figure 6a-d. F4, example of gesture suspension and gesture phase labeling. In a., the hands hold with negligible movement until b., the hands start up in a preparation phase (here, the middle of the movement is shown). In c., the hands move swiftly downward in a stroke phase (the middle of the movement is shown). Finally, in d., the hands hold above rest position (hands at sides) before continuing onto the following movement.

4. Results and analysis

4.1 Region durations

For each subject, the following durations are listed in Table 1 below: total speech excerpt duration, total pause duration (the sum of all pause durations within the speech excerpt), total fluent area labeled, total pre-pause region duration, and total post-pause region duration.

Subject	Speech Excerpt	Total Pause Duration	Total Fluent Area Labeled	Total Pre-Pause	Total Post-Pause
F2	68.318	10.349	10.501	10.405	8.301
F3	31.532	5.457	5.483	3.919	2.527
F4	78.678	23.235	20.324	13.108	8.031
F5	79.896	16.572	17.976	13.436	7.715
M1	52.266	7.984	9.993	6.994	6.017
M2	145.078	33.995	35.313	15.87	9.93

Table 1. Speech region durations for all subjects

Speech durations varied across subjects from a minimum of 31.532 seconds (F3) to a maximum of 145.078 seconds (M2). In each subject, pause durations were approximately one-fifth the duration of the corresponding speech excerpt. The total amount of fluent area labeled for a subject was approximately matched with the subject's corresponding total pause duration, so that similar amounts of each speech region could be observed. Pre-pause and post-pause region durations were not controllable, as they depended on the length of the words uttered just before and after each pause.

4.2 Raw gesture data

For each subject, the following raw gesture data is listed below: in Table 2, onset frequency occurring in pauses, fluent speech (i.e. fluent area labeled), the pre-pause region, and the post-pause region; and in Table 3, target frequency occurring in the same speech regions respectively.

Subject	Onset Frequency			
	Pause	Fluent Speech	Pre-Pause	Post-Pause
F2	11	14	8	15
F3	1	6	3	4
F4	20	18	7	9
F5	12	22	15	15
M1	14	13	9	6
M2	16	29	13	8

Table 2. Gesture onset frequency data for all subjects

Subject	Target Frequency			
	Pause	Fluent Speech	Pre-Pause	Post-Pause
F2	9	12	12	13
F3	1	7	4	2
F4	16	18	6	13
F5	12	21	10	12
M1	7	12	7	14
M2	16	30	16	17

Table 3. Gesture target frequency data for all subjects

4.3 Gesture frequency and rate

4.3.1 Gesture onset frequency and rate: in pause vs. fluent speech

Gesture onset frequency, or the total number of gesture onsets, is represented below via bar graph. Gesture onset rate, or the average number of gesture onsets per second, is also represented below, to the right of onset frequency graphs. Gesture onset rate was calculated by taking the subject's onset frequency in a speech region and dividing by the total duration for that speech region.

Because gesture suspensions refer to the absence of gesture movement between gesture movements, suspensions were not included in either calculation.

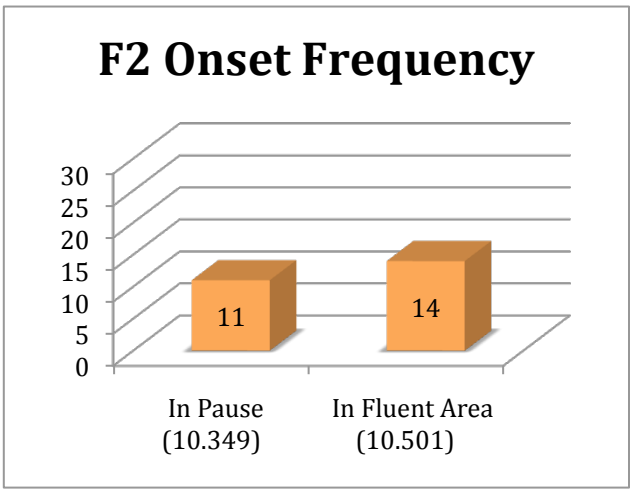


Figure 7. F2 onset frequency

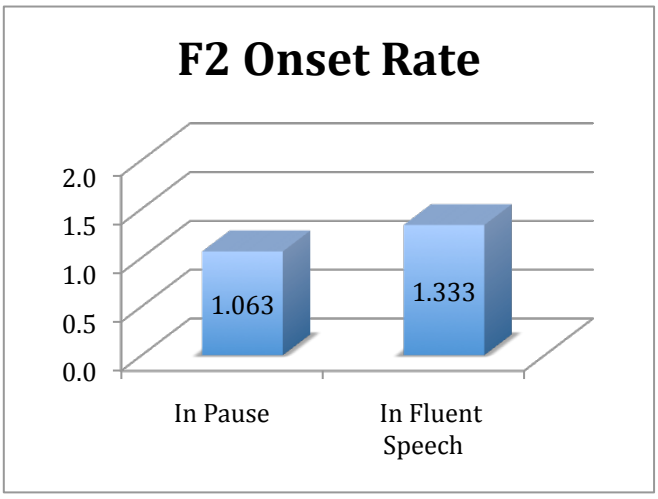


Figure 8. F2 average onset per second

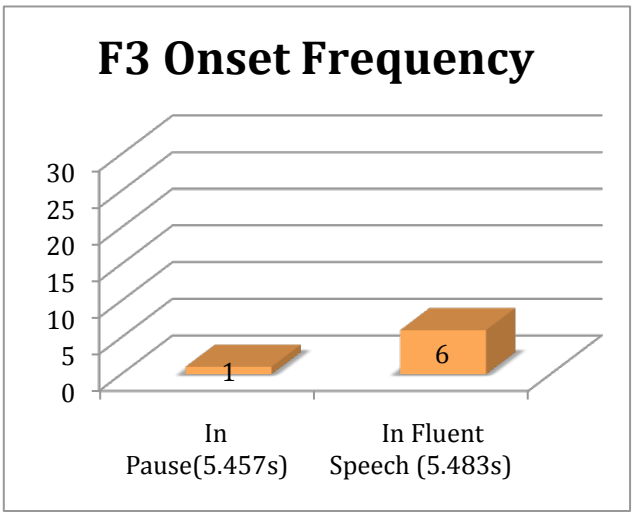


Figure 9. F3 onset frequency

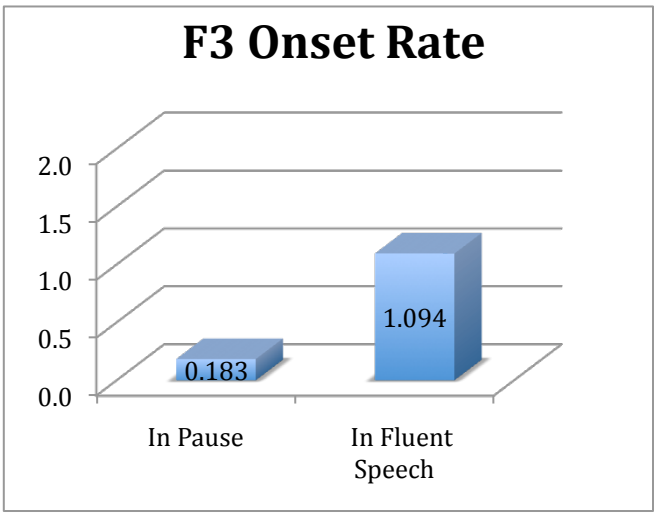


Figure 10. F3 average onset per second

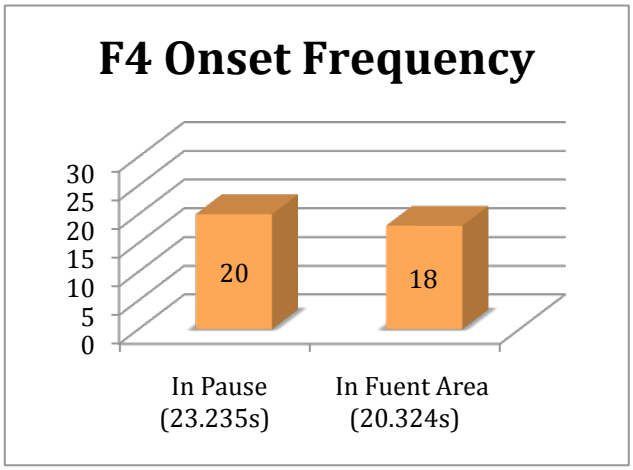


Figure 11. F4 onset frequency

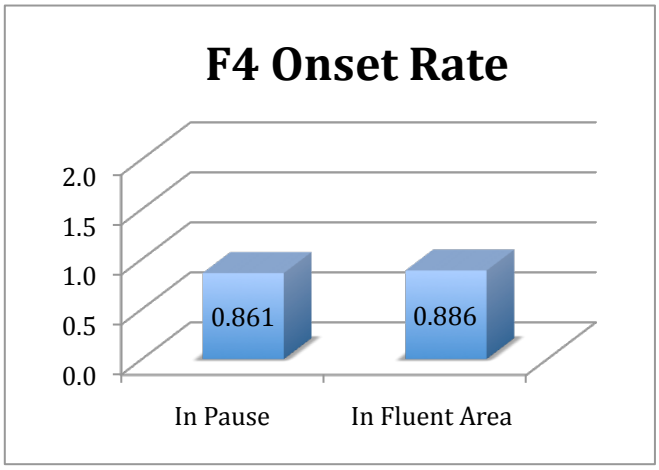


Figure 12. F4 average onset per second

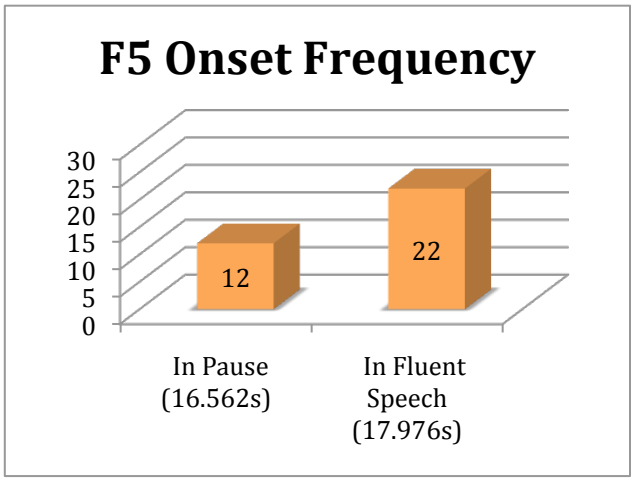


Figure 13. F5 onset frequency

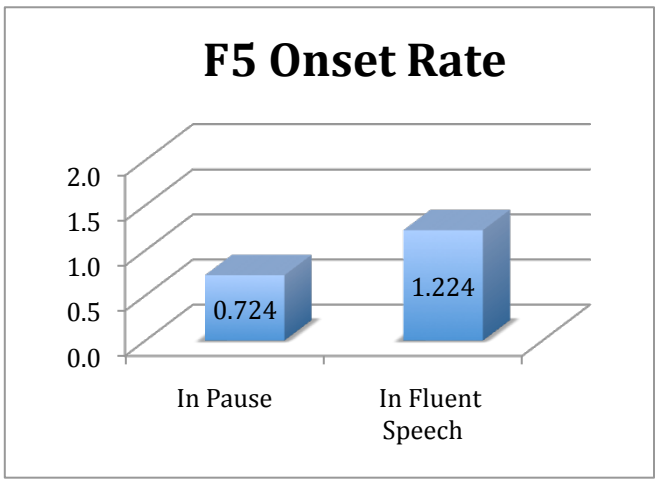


Figure 14. F5 average onset per second

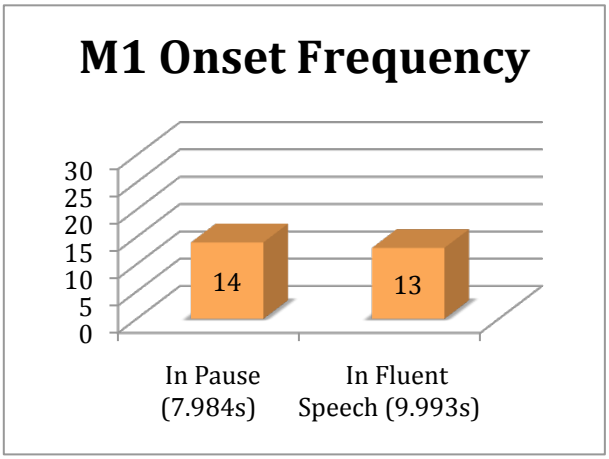


Figure 15. M1 onset frequency

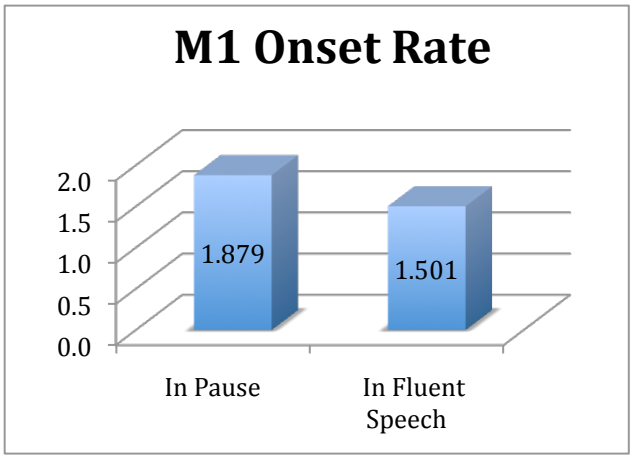


Figure 16. M1 average onset per second

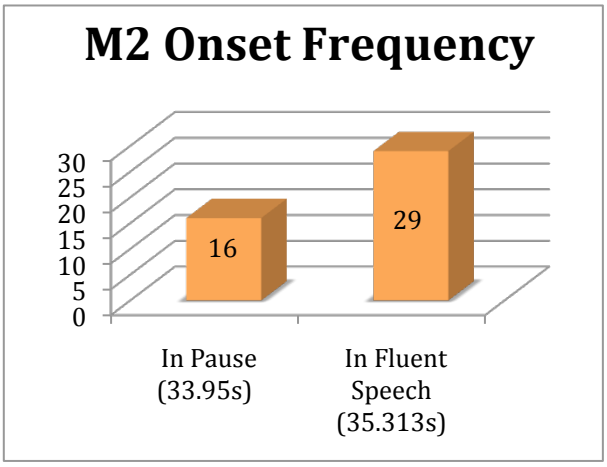


Figure 17. M2 onset frequency

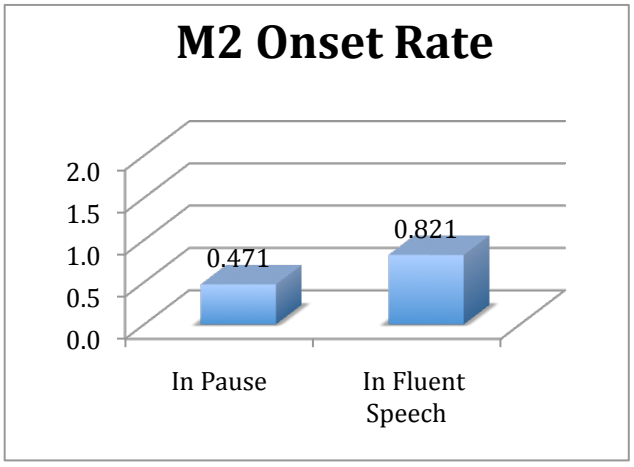


Figure 18. M2 average onset per second

Results indicate that for most subjects, onsets occur at a slightly lower rate in pauses than in fluent speech. For subject M2, onsets occurred at a much lower rate in pauses than in fluent speech (approximately half as often). F4 and M1 show a different pattern; however, given the subjects' frequency values only differ by one to two onsets, this does not indicate that for these subjects onsets occur more often in pauses than in fluent speech, but rather that onsets occur at about the same frequency in pauses and fluent speech.

These results indicate that speakers are equally likely or slightly less likely to begin gestures (of any phase: preparation, stroke, or release) in pauses than in fluent speech.

4.3.2 Gesture onset frequency and rate: pre- and post-pause regions

Onset rates were also calculated for pre- and post-pause regions, in order to determine whether differences in gesture onset exist near pause boundaries. Frequencies are not represented via bar graph, as pre- and post-pause durations vary from pause and fluent speech durations (see Table 1 above; see Table 2 for onset frequencies in pre- and post-pause regions). Pause and fluent speech rates are included for comparison.

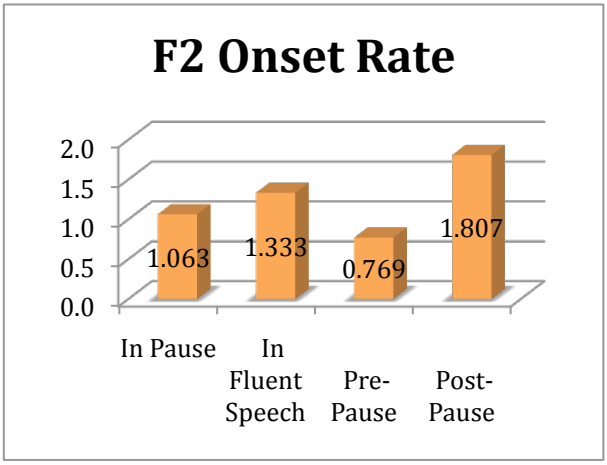


Figure 19. F2 average onset per second

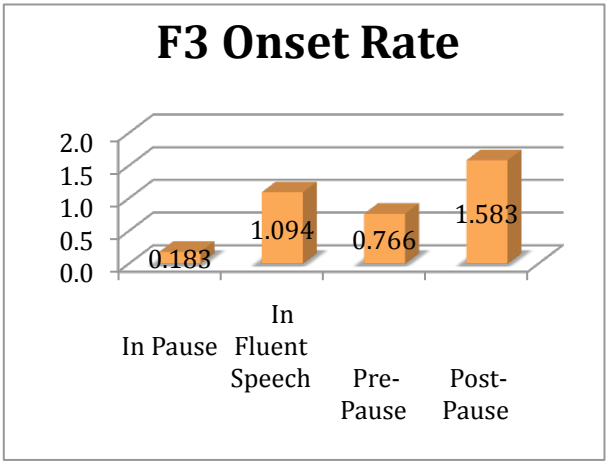


Figure 20. F3 average onset per second

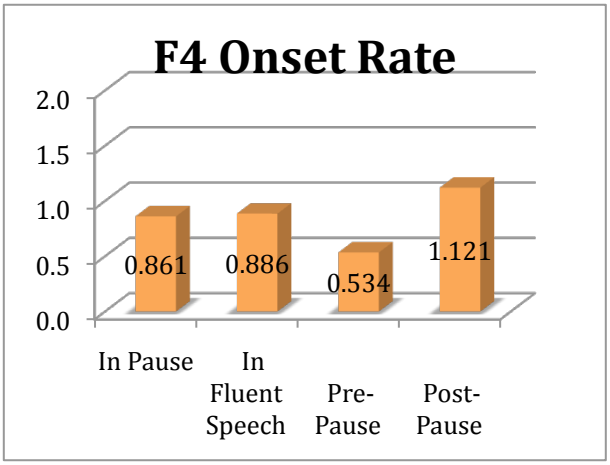


Figure 21. F4 average onset per second

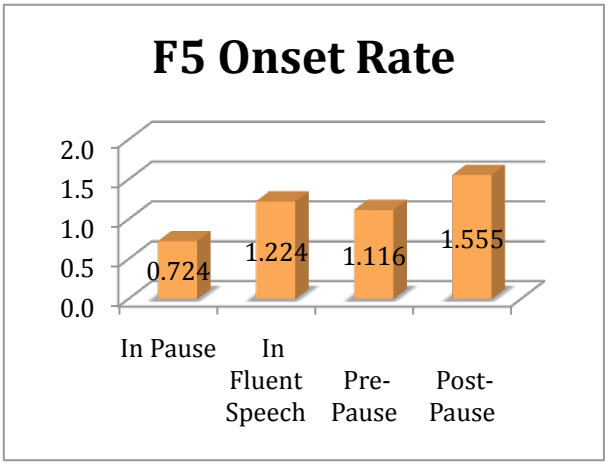


Figure 22. F5 average onset per second

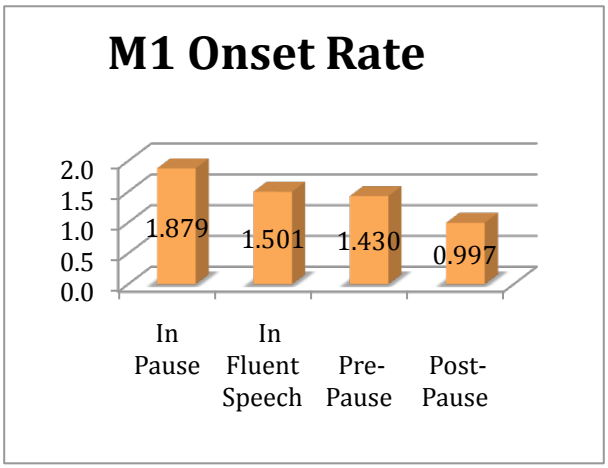


Figure 23. M1 average onset per second

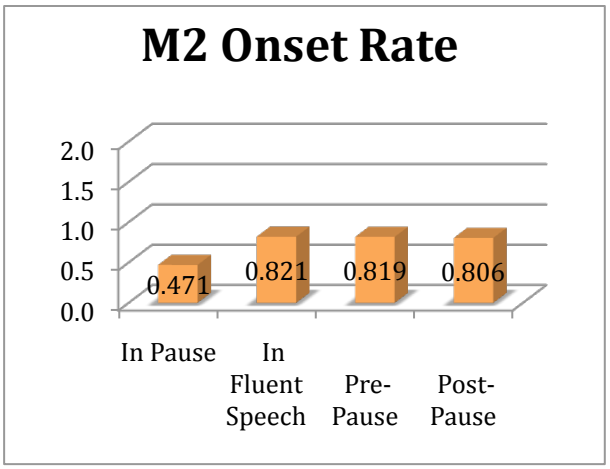


Figure 24. M2 average onset per second

Results indicate that for most subjects, onsets occur at a slightly lower rate in the pre-pause region than in fluent speech. This implies that speakers are less likely to begin gesturing just before a pause than they are to begin gesturing in regular speech with no upcoming pause. Conversely, in the post-pause region, for most subjects, onsets occur at a slightly higher rate than in any other speech regions. This indicates that speakers are more likely to begin gesturing immediately following a pause than anywhere else in speech.

Subject M1 displayed different post-pause behavior from the other subjects, with onsets occurring at the lowest rate in the post-pause region out of all speech regions. Subject M2's onset rates were distributed fairly equally across fluent speech, the pre-pause region, and the post-pause region.

4.3.3 Gesture target frequency and rate: in pause vs. fluent speech

Gesture target frequencies, or the total number of gesture targets, are represented below via bar graph, as well as gesture target rate. Average target rate for each subject was calculated by taking the subject's target frequency in a speech region and dividing by the total duration for that speech region, resulting in an average gesture onset per second. Once again, gesture suspensions were not included in either calculation.

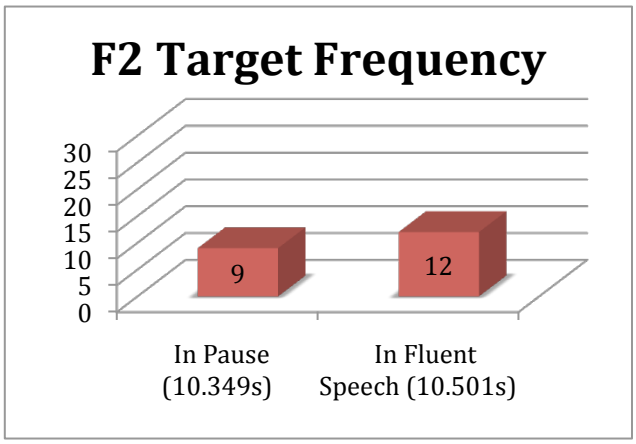


Figure 25. F2 target frequency

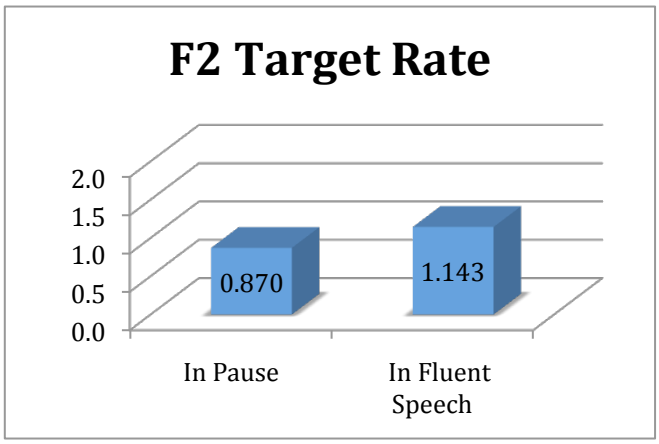


Figure 26. F2 average target per second

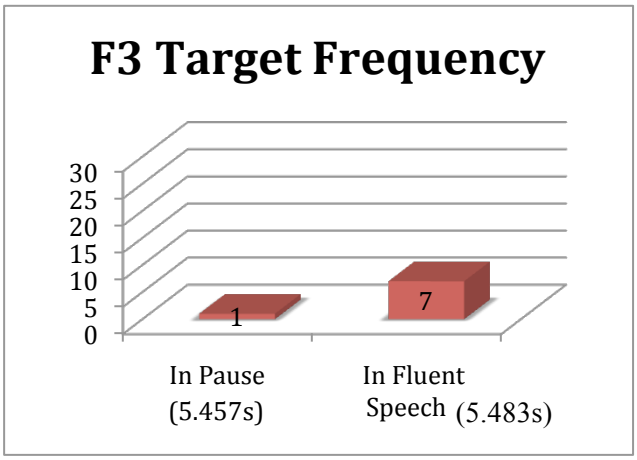


Figure 27. F3 target frequency

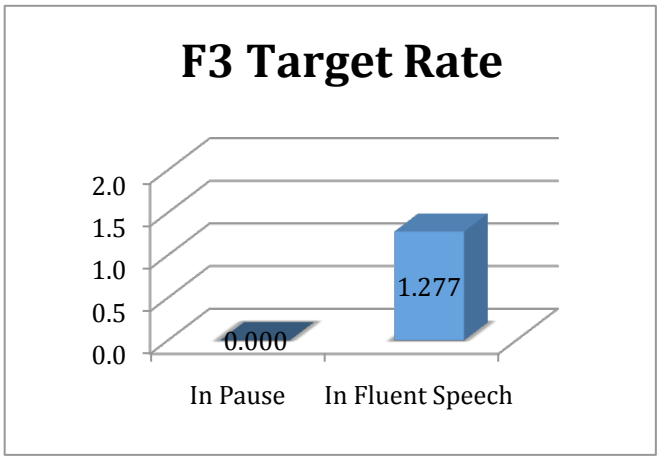


Figure 28. F3 average target per second

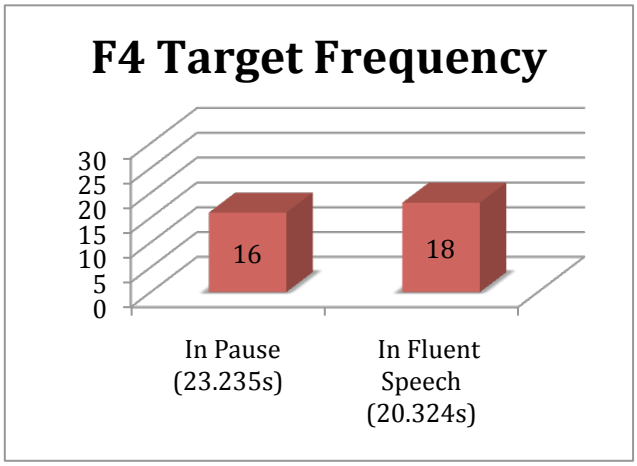


Figure 29. F4 target frequency

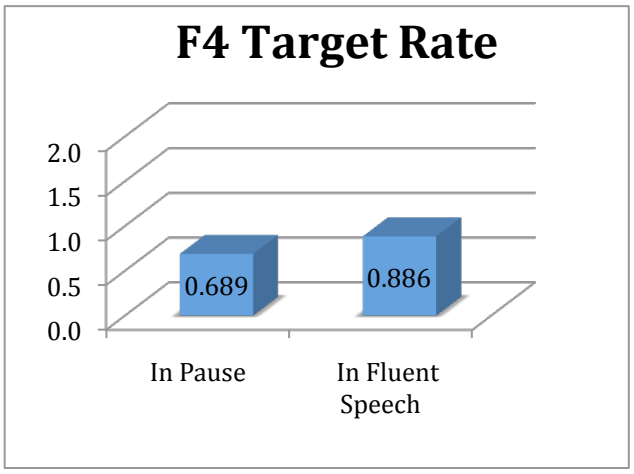


Figure 30. F4 average target per second

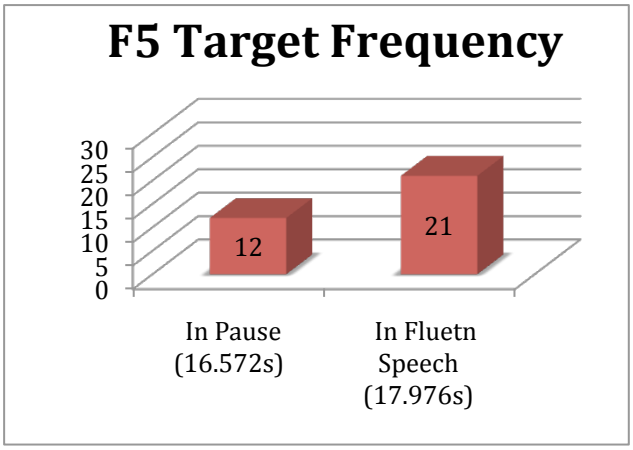


Figure 31. F5 target frequency

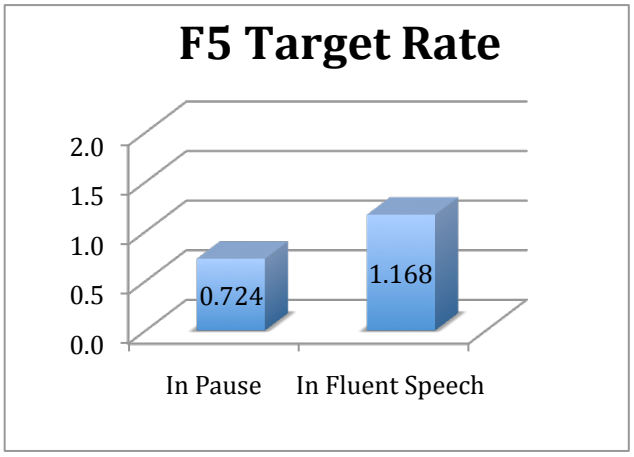


Figure 32. F5 average target per second

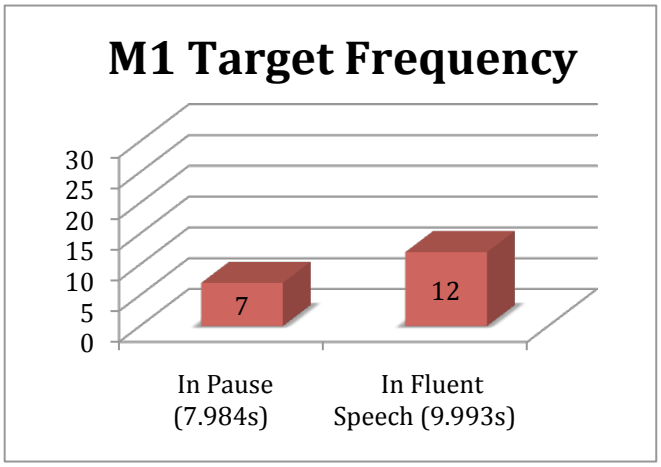


Figure 33. M1 target frequency

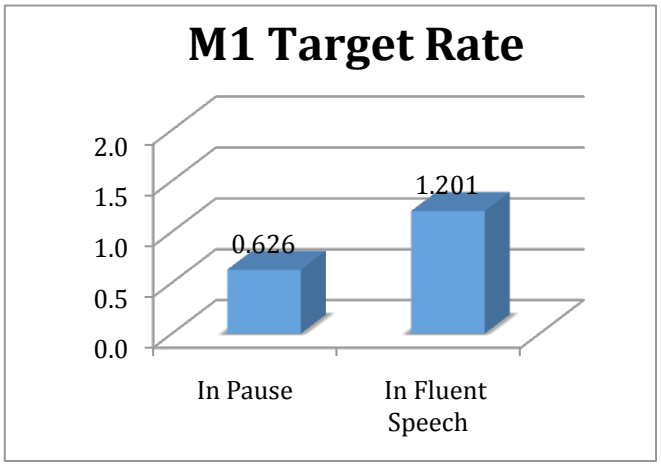


Figure 34. M1 average target per second

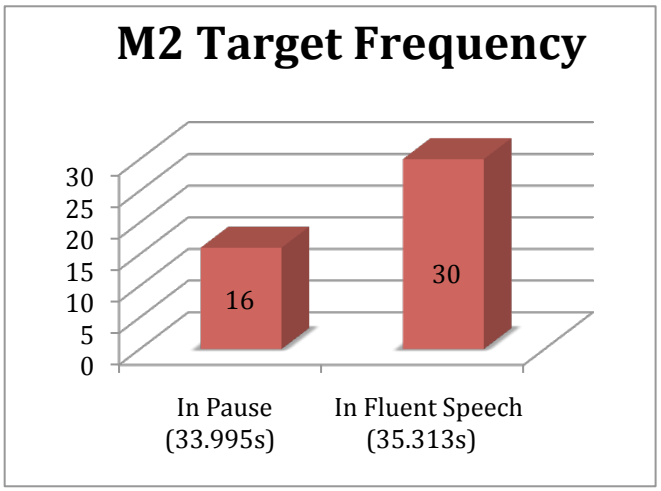


Figure 35. M2 target frequency

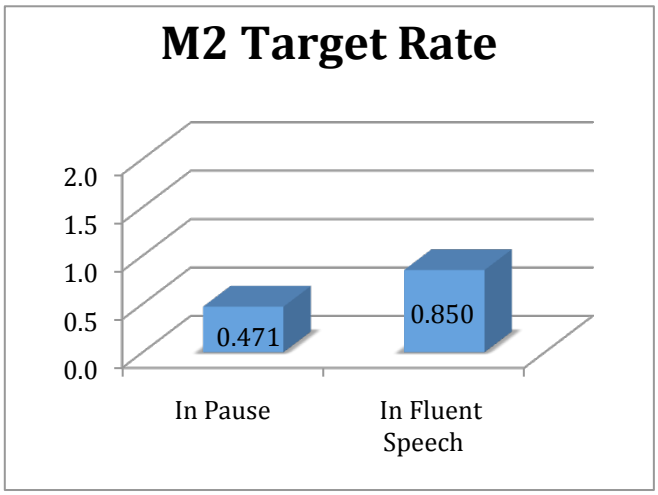


Figure 36. M2 average target per second

Results indicate that for all subjects, targets occur at a slightly lower rate in pauses than in fluent speech. This implies that speakers are less likely to complete gestures (of any phase: preparation, stroke, or release) in pauses than in fluent speech.

4.3.4 Gesture target frequency and rate: pre- and post-pause regions

Target rates were also calculated for pre- and post-pause regions, in order to determine whether differences in gesture target exist near pause boundaries. As with the onset frequencies above, target frequencies are not represented via bar graph, as pre- and

post-pause durations vary from pause and fluent speech durations (see Table 1 above; see Table 3 for target frequencies in pre- and post-pause regions). Pause and fluent speech rates are included for comparison.

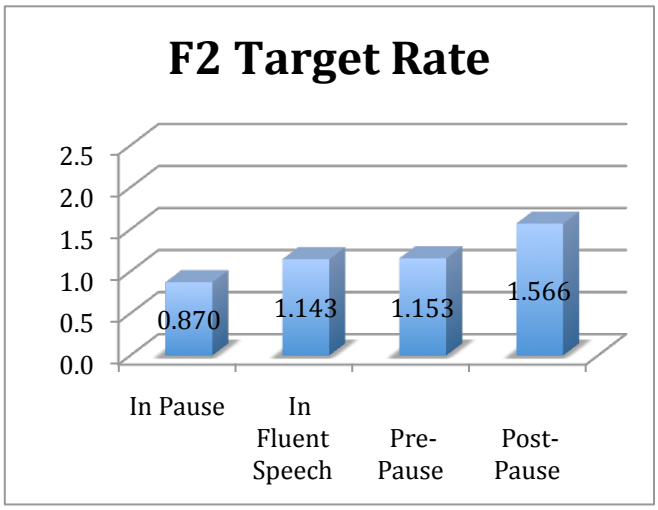


Figure 37. F2 average target per second

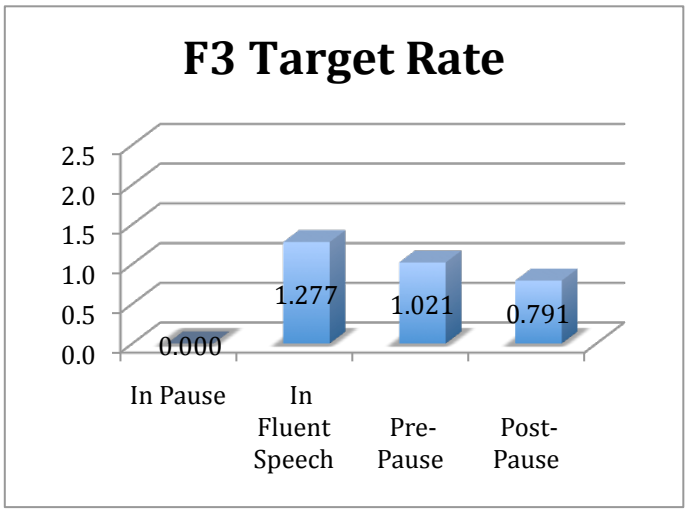


Figure 38. F3 average target per second

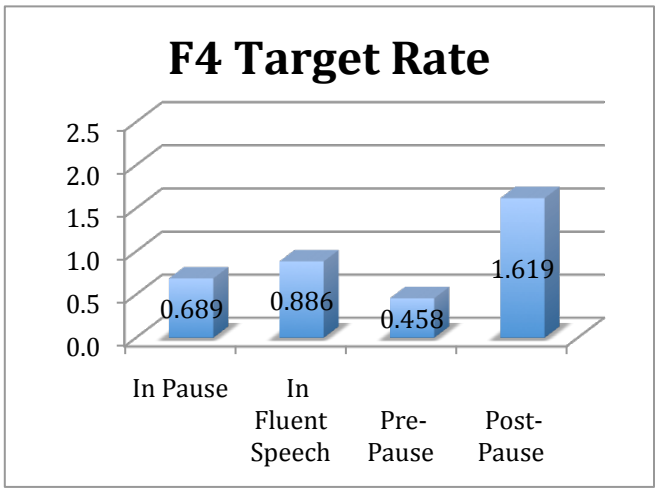


Figure 39. F4 average target per second

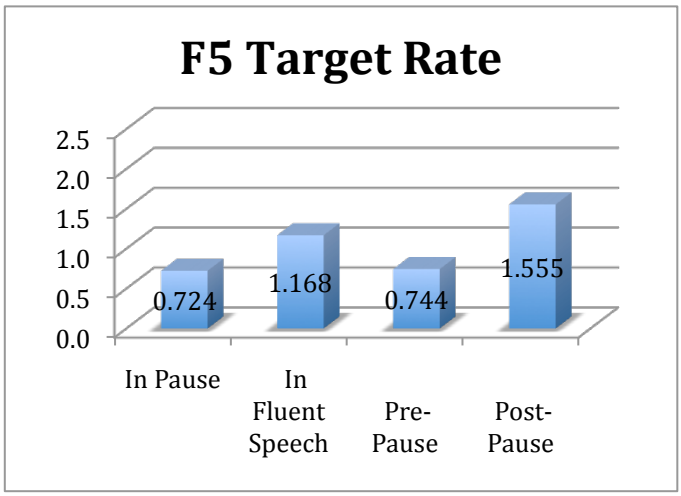


Figure 40. F5 average target per second

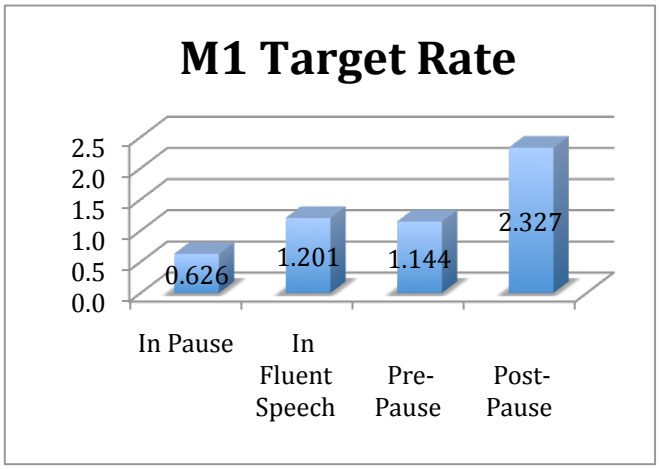


Figure 41. M1 average target per second

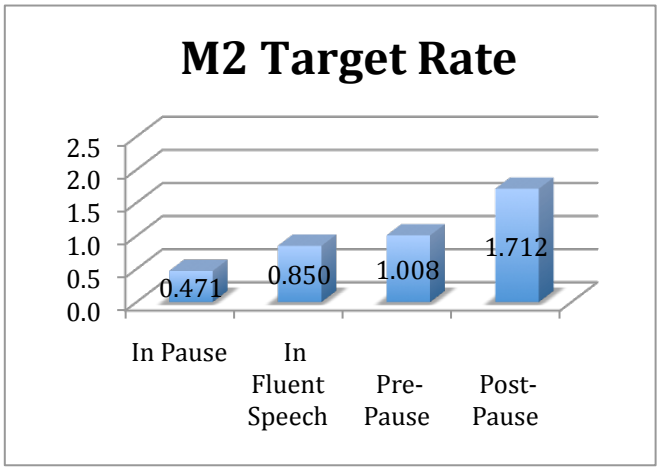


Figure 42. M2 average target per second

Results indicate that targets occur slightly less frequently in the pre-pause region than in fluent speech. This implies that speakers are less likely to complete gesturing just before a pause than they are to complete gesturing in regular speech with no upcoming pause. Conversely, in the post-pause region, for most subjects, targets occur at a slightly higher rate than in other speech regions. This indicates that speakers are more likely to complete a gesture immediately following a pause than anywhere else in speech.

4.3.5 Gesture suspension frequency and rate: all four speech regions

Suspension frequencies and rates were examined in order to determine if any pattern would emerge between suspension occurrence and speech region. Suspension frequencies show raw suspension onset and target data for each speech region. Suspension onset and target rates were calculated by taking the subject's suspension onset or target frequency in a speech region and dividing by the total duration for that speech region, resulting in an average gesture suspension onset or target per second. Suspension rate graphs are below suspension frequency graphs.

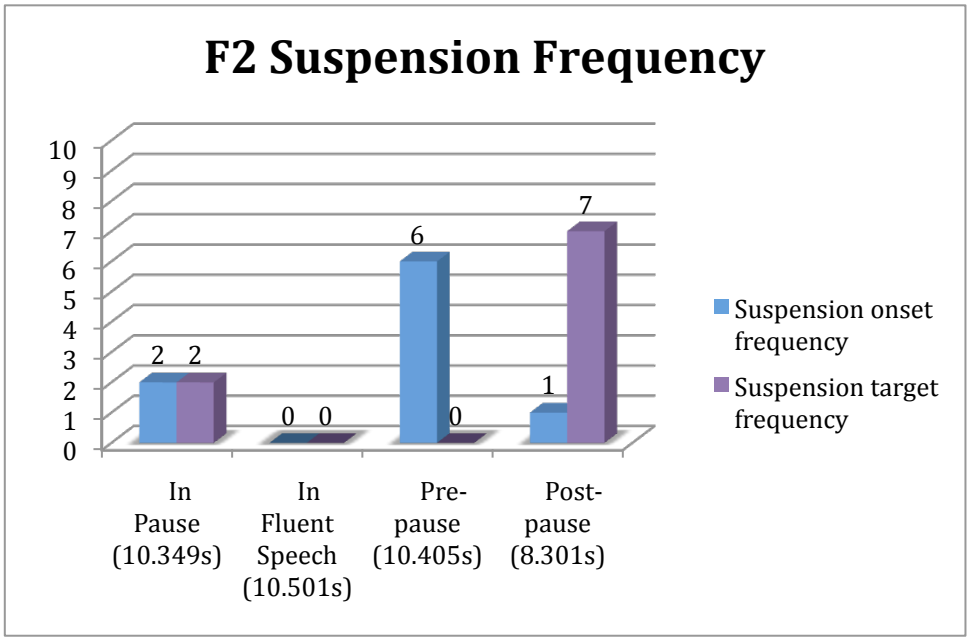


Figure 43. F2 suspension frequency

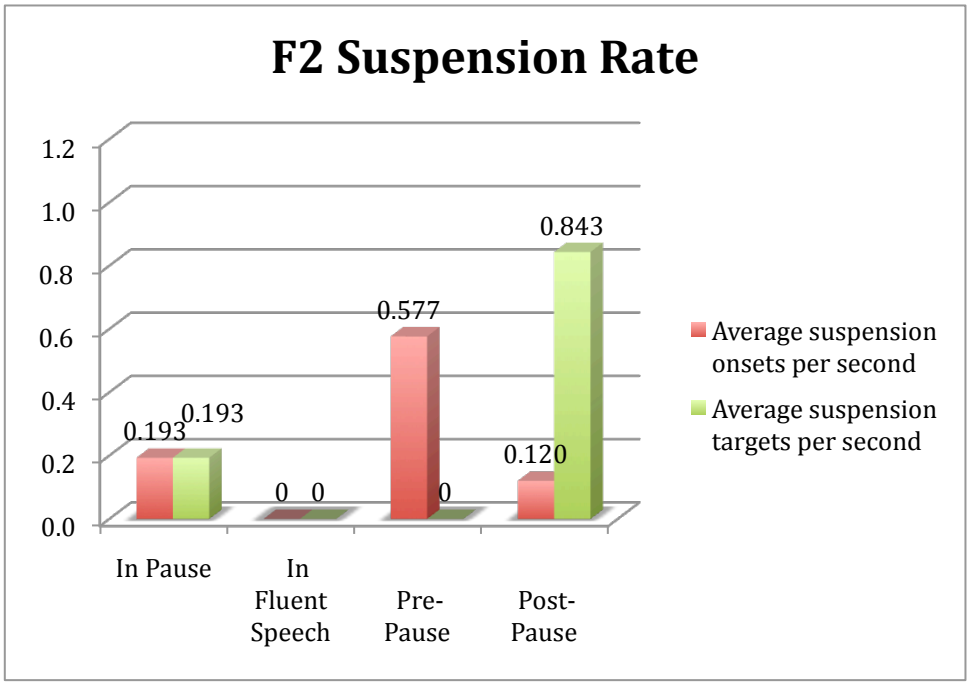


Figure 44. F2 average suspension onset and target per second

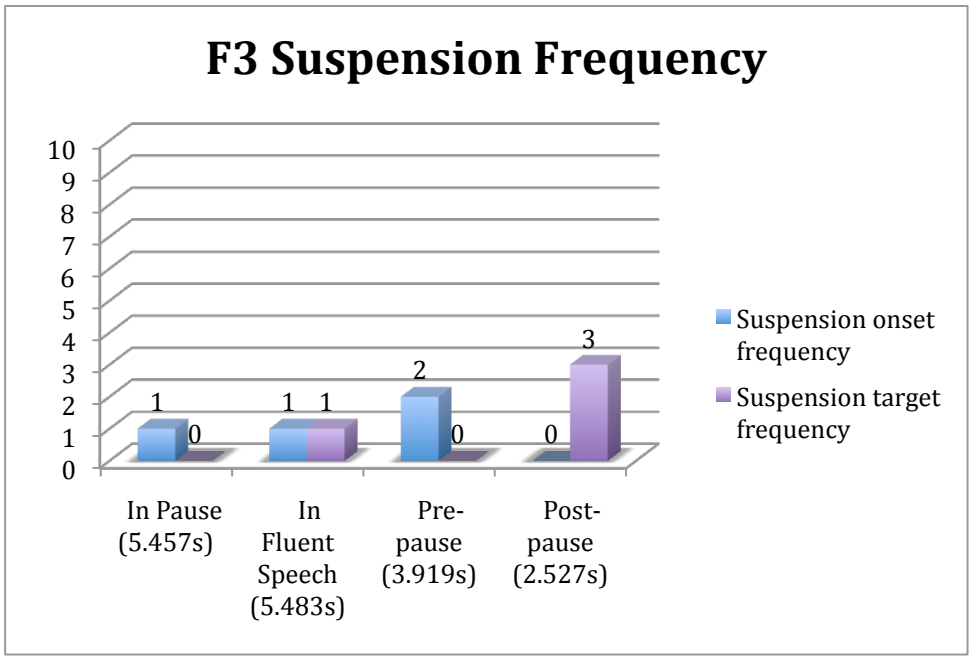


Figure 45. F3 suspension frequency

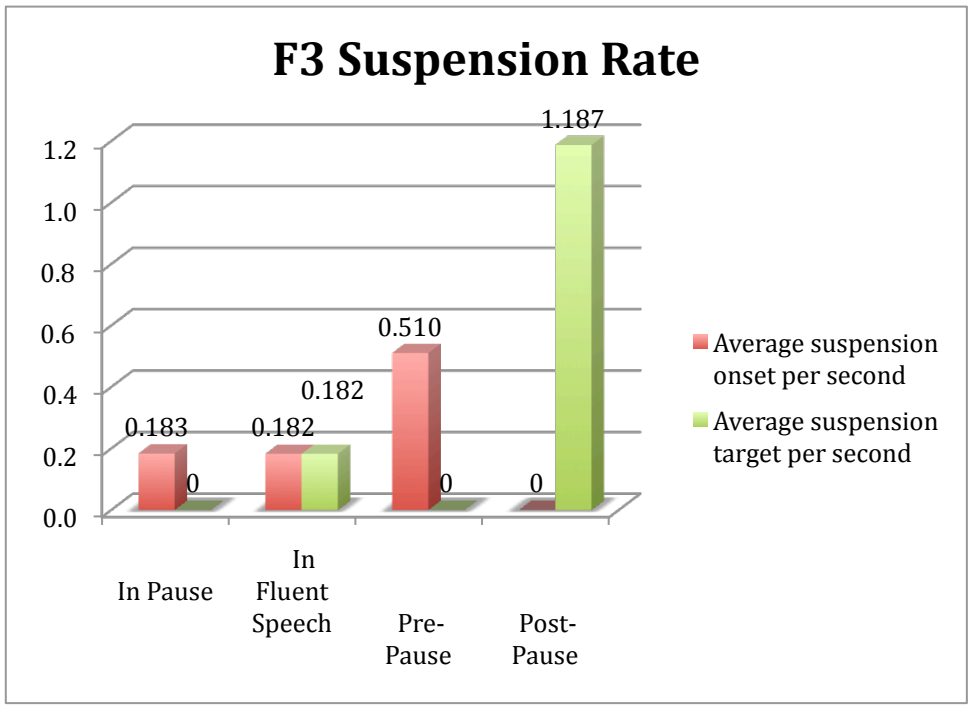


Figure 46. F3 average suspension onset and target per second

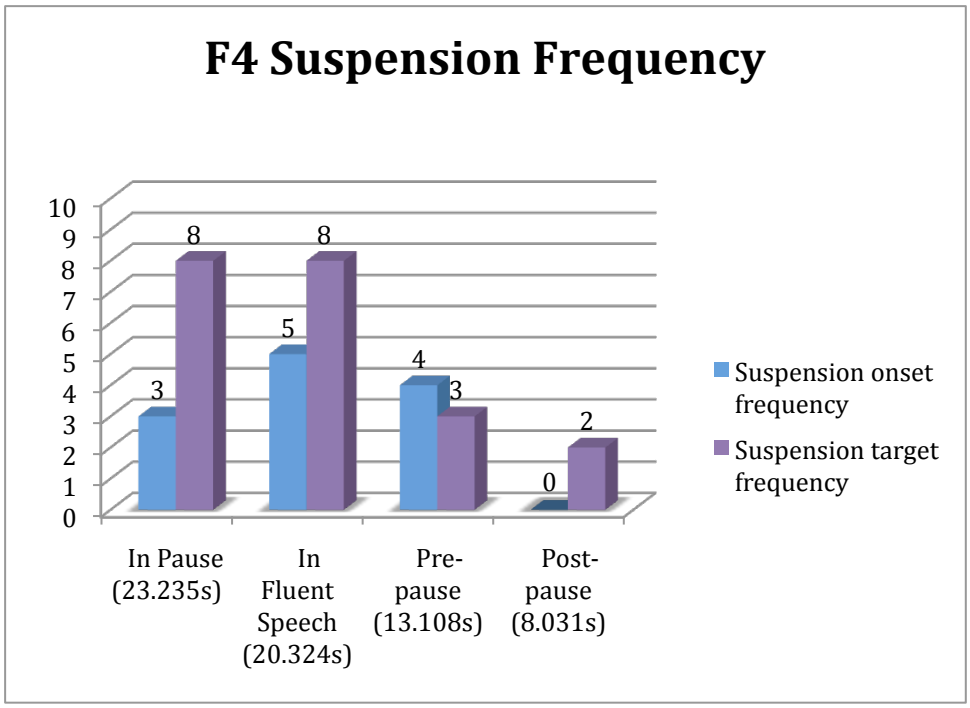


Figure 47. F4 suspension frequency

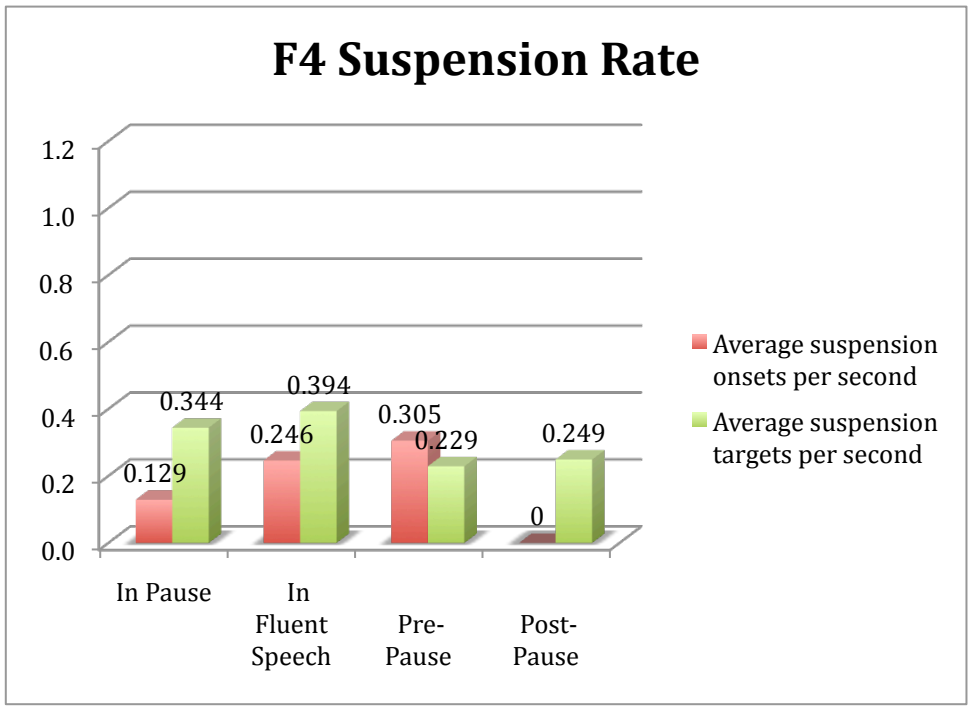


Figure 48. F4 average suspension onset and target per second

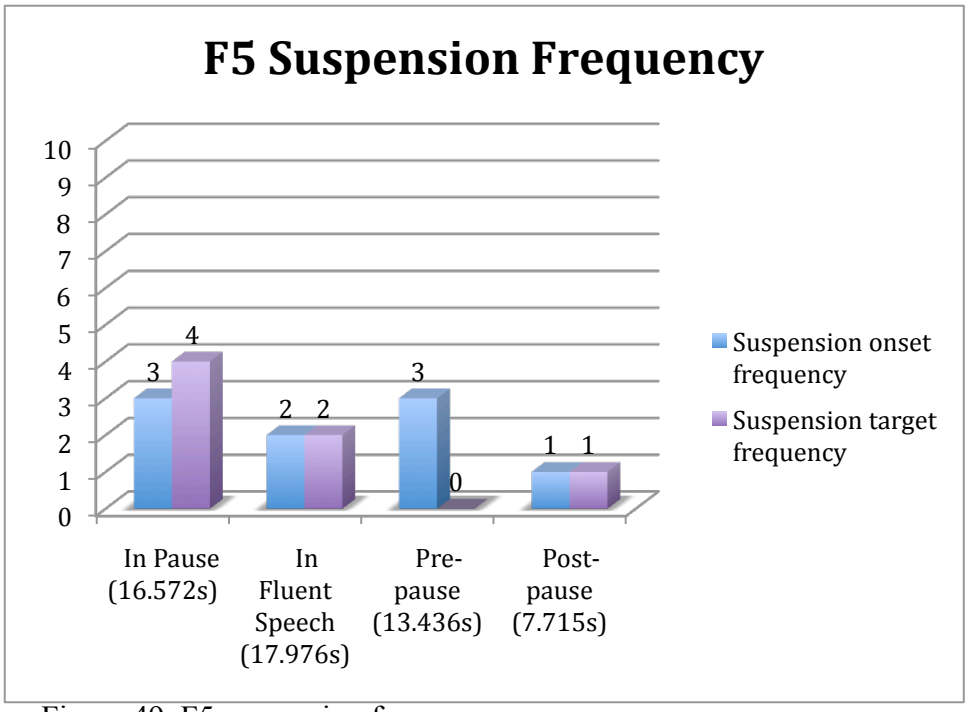


Figure 49. F5 suspension frequency

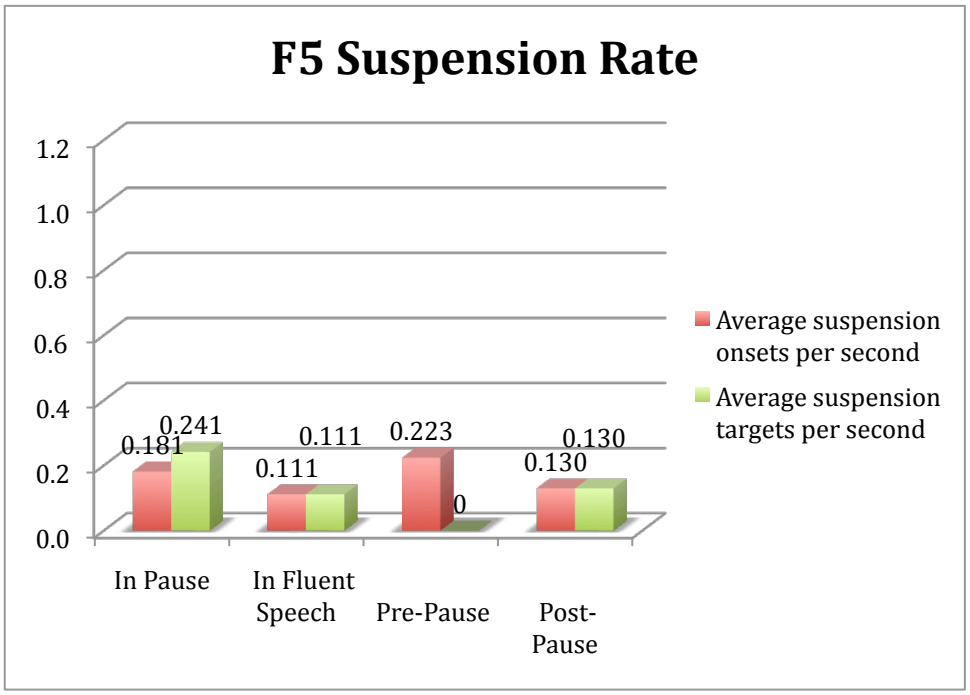


Figure 50. F5 average suspension onset and target per second

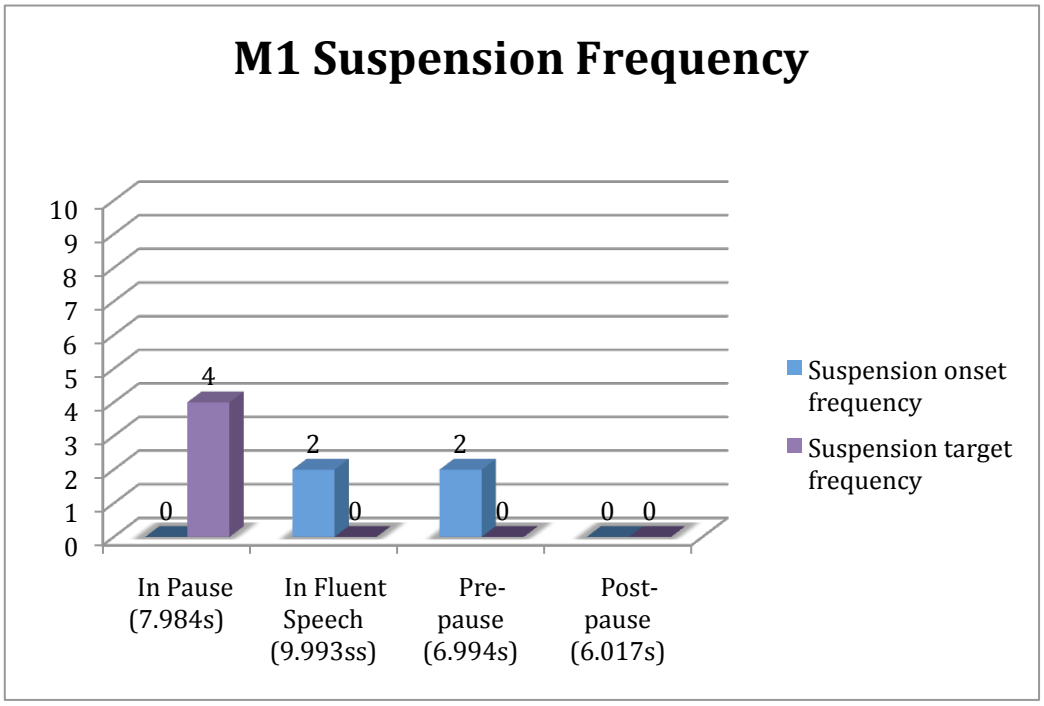


Figure 51. M1 suspension frequency

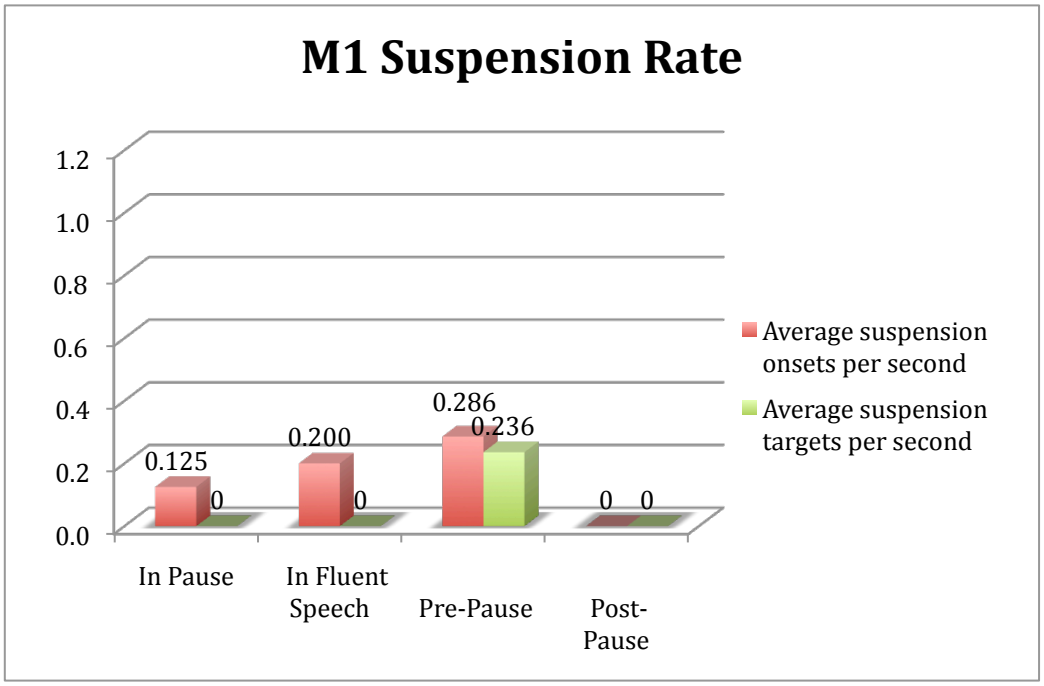


Figure 52. M1 average suspension onset and target per second

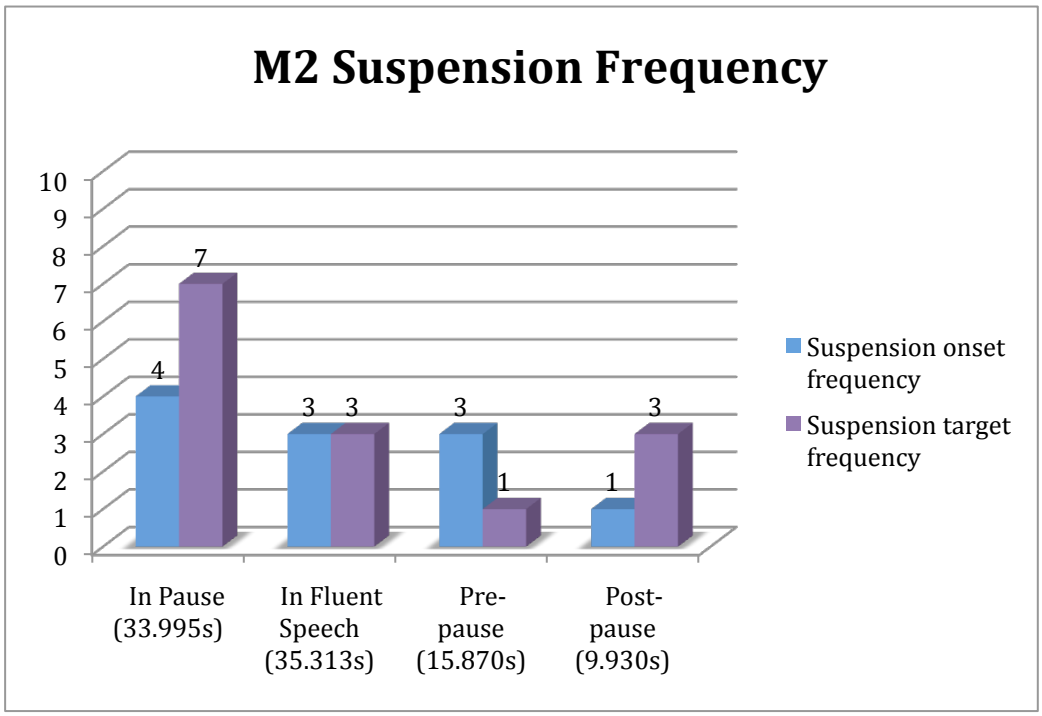


Figure 53. M2 suspension frequency

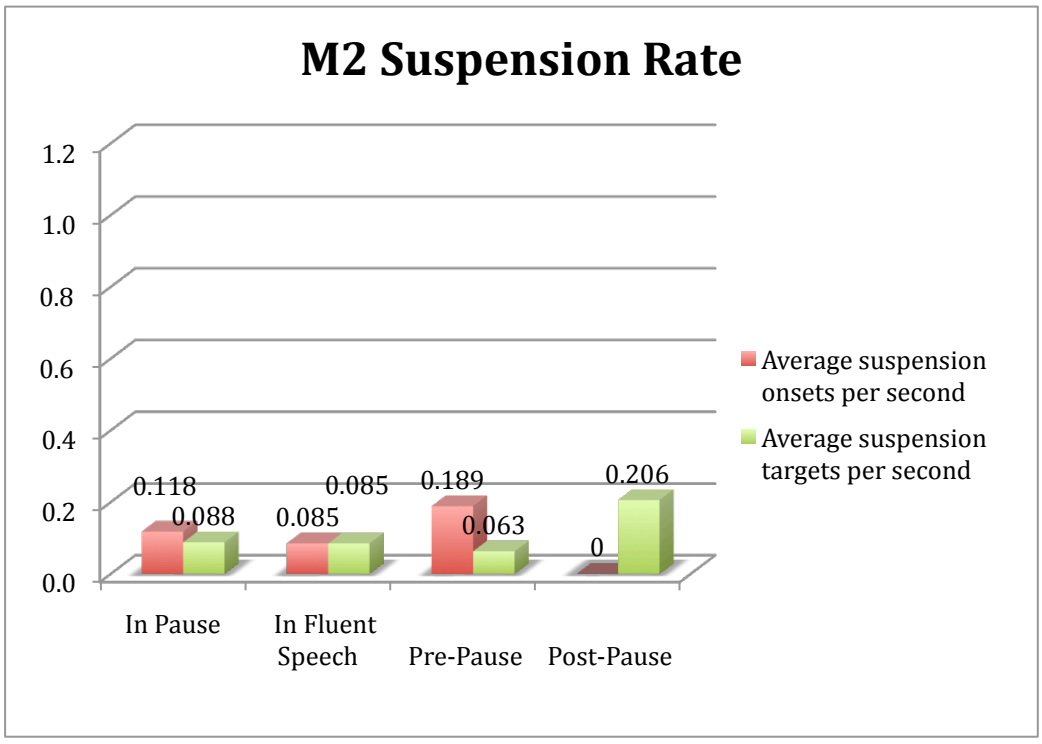


Figure 54. M2 average suspension onset and target per second

Overall, gesture suspension onsets and targets were more prevalent in pauses and pause regions, although they did also occur in fluent speech for some subjects, especially the two hand gesturers, F4 and M2. Onsets consistently occurred at a higher rate in the pre-pause region than in any other speech region, while targets were distributed differently for different subjects.

4.4 Complete gestures

4.4.1 Complete gestures: in pause vs. in fluent speech

Perhaps the most important calculation is complete gesture location. When a gesture starts in a pause (its onset occurs in a pause), it could also end in the pause, or it could end after the pause, i.e. in the post-pause region or in following fluent speech. Similarly, when a gesture ends in a pause, the gesture could also have started in the pause, or it could have started before the pause, i.e. in the pre-pause region or in previous fluent speech. Thus, it is important to examine not only the frequency of onsets and targets separately, but also the location of complete onset-target gesture units in pauses and in fluent speech.

Complete gestures were calculated by measuring the frequency of gestures with both onset and target occurring in the same speech region. Because marked off fluent areas were dispersed across speech and chunked similarly to pauses, they possessed approximately equal likelihood of having complete gestures within them or cut-off gestures at their edges.

Areas of speech not marked off for any speech regions were labeled "Unmarked." These "Unmarked" areas represent the remaining speech not marked off for labeling in comparison with pauses. Therefore, gestures occurring at fluent area edges could simply be continuing into additional fluent speech, i.e. an "Unmarked" area. However, in order to maintain similar durations between fluent areas and pauses, these fluent area edge gestures were not taken into account for complete gesture calculations.

Results are detailed by subject in Table 4 below. F3 was omitted from complete gesture calculations due to the already extremely low frequency of gesture onsets and targets in pauses (see Tables 2 and 3).

Subject	Speech Region	Onsets	Targets	Complete Gestures Possible	Complete Gestures Actual	Actual/Possible (%)
F2	In Pause	11	9	9	4	44.4
	Fluent Area	14	12	12	11	91.7
F4	In Pause	20	16	16	5	31.3
	Fluent Area	18	18	18	12	66.7
F5	In Pause	12	12	12	5	41.7
	Fluent Area	22	21	21	16	76.2
M1	In Pause	14	7	7	4	57.1
	Fluent Area	13	12	12	11	91.7
M2	In Pause	16	16	16	4	25
	Fluent Area	29	30	29	24	82.8

Table 4. Complete gestures in pauses vs. in fluent speech

These results indicate that despite a prevalence of onsets and targets in pauses, complete onset-target gesture units occur far less frequently in pauses than in fluent speech. This implies that gestures more often bridge in or out of the pause than occur completely inside of it.

4.4.2 Complete gesture locations: behavior at pause areas

Because a gesture could begin within the pause or a pre- or post-pause region and end in an "Unmarked" region, or vice versa, the "Unmarked" category was included in calculations for complete gestures surrounding pauses (see Section 3.2.5.1 for discussion of speech area labels). This section focuses on complete gesture behavior at pauses and does not compare with fluent speech, thus allowing the inclusion of "Unmarked" areas.

In order to determine complete gesture behavior in and around pauses, gesture onset-target units that occurred in a pause area were measured for their specific location in relation to the pause, and categorized as "in pause," "bridge in," "bridge out," or "pre to post." These locations were labeled according to the following criteria:

- In Pause: gesture onset and target both occur within the pause
- Bridge IN: gesture onset occurs in pre-pause region, fluent area, or unmarked area, gesture target occurs in pause
- Bridge OUT: gesture onset occurs in pause, gesture target occurs in post-pause region, fluent area, or unmarked area
- Pre to Post: gesture onset occurs in pre-pause region, gesture target occurs post-pause region (pause is contained within gesture)

Complete gesture locations are given for each subject below.

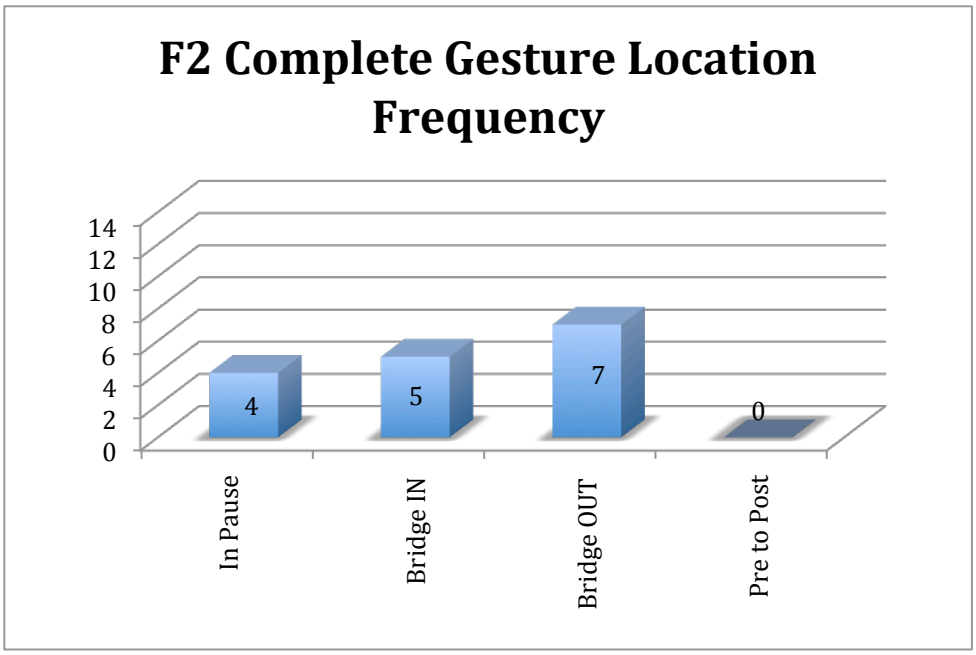


Figure 55. F2 complete gesture location frequency

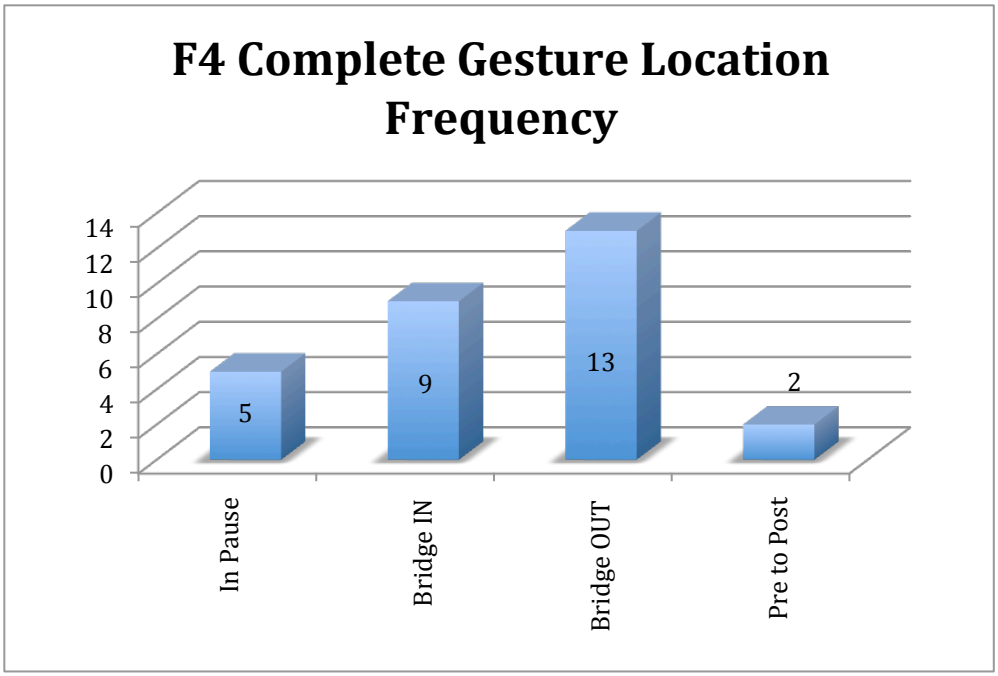


Figure 56. F4 complete gesture location frequency

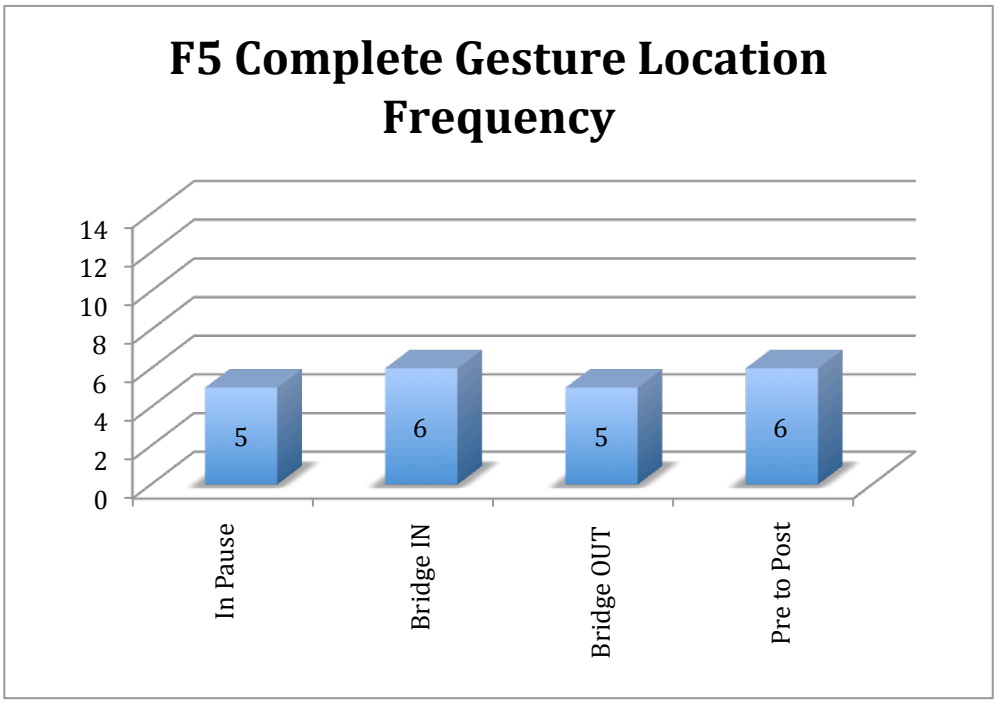


Figure 57. F5 complete gesture location frequency

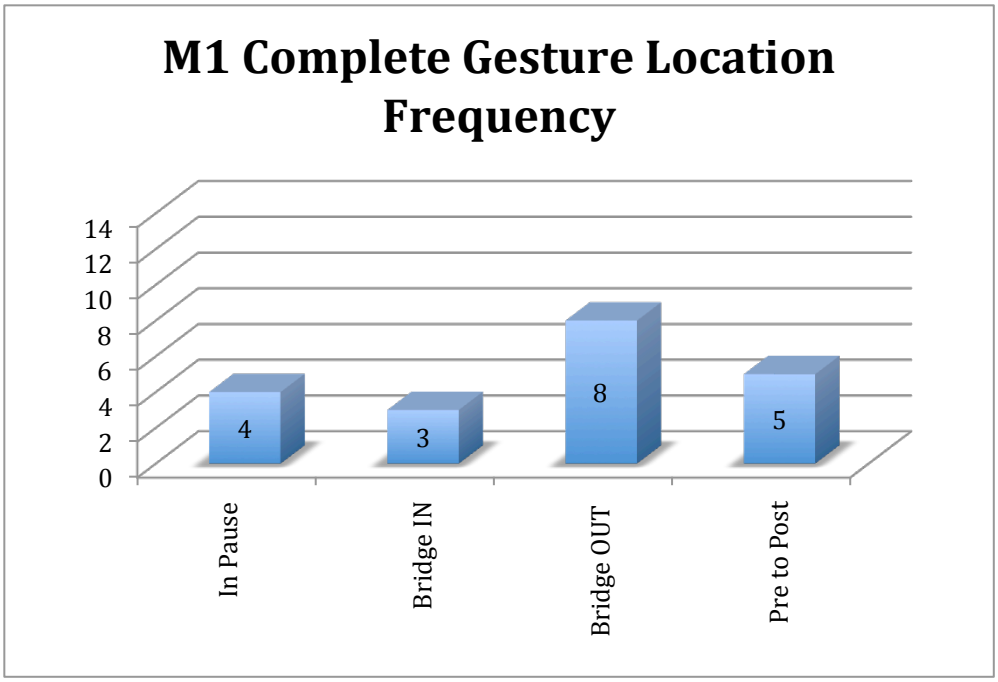


Figure 58. M1 complete gesture location frequency

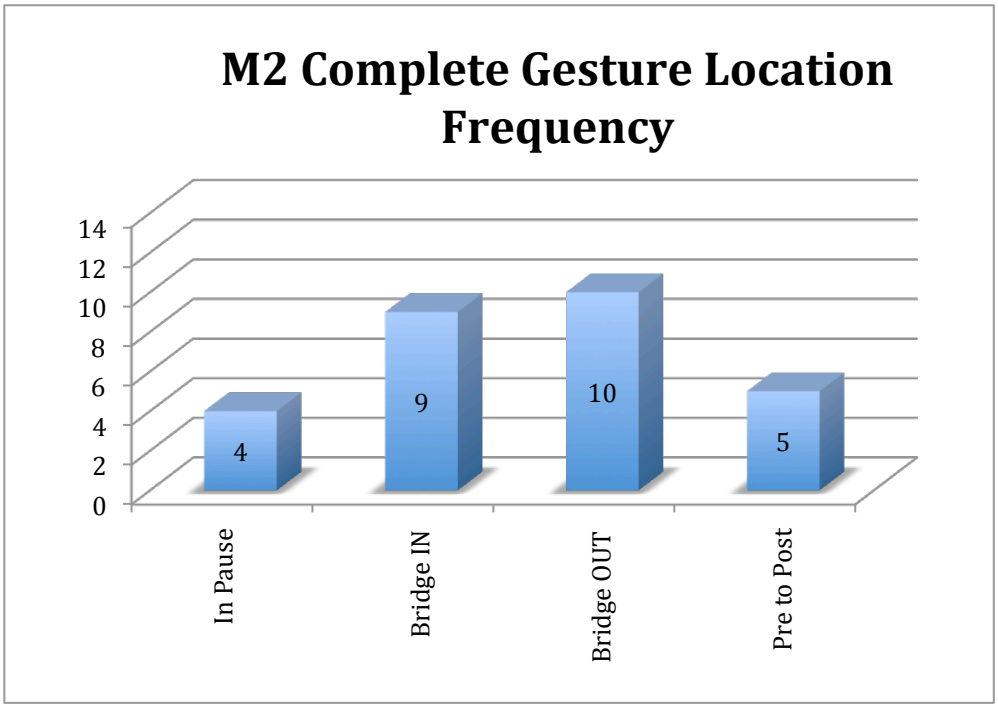


Figure 59. M2 complete gesture location frequency

Results indicate that for all subjects but F5, complete gestures occur most often bridging out of the pause. This corroborates the result that the post-pause region contains

a higher number of targets. Subject F5's complete gestures were distributed fairly evenly across the four locations.

4.4.3 Complete suspension locations: behavior at pause areas

Complete gesture suspension locations were also calculated, in order to determine the distribution of gesture suspensions in comparison with the distribution of regular gestures.

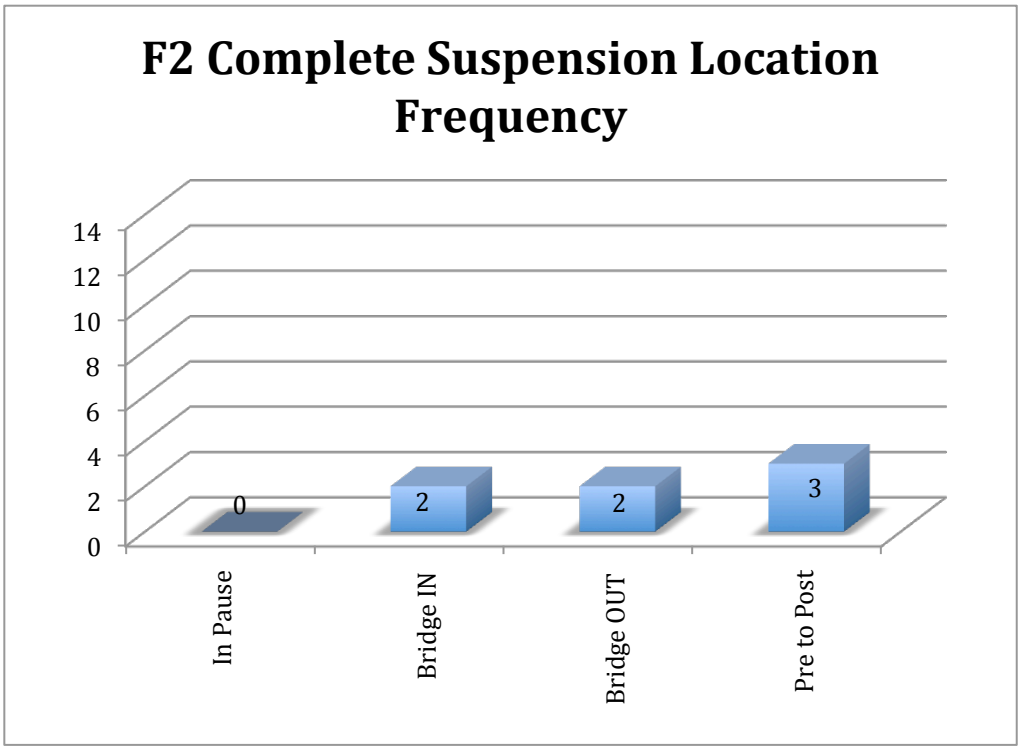


Figure 60. F2 complete suspension location frequency

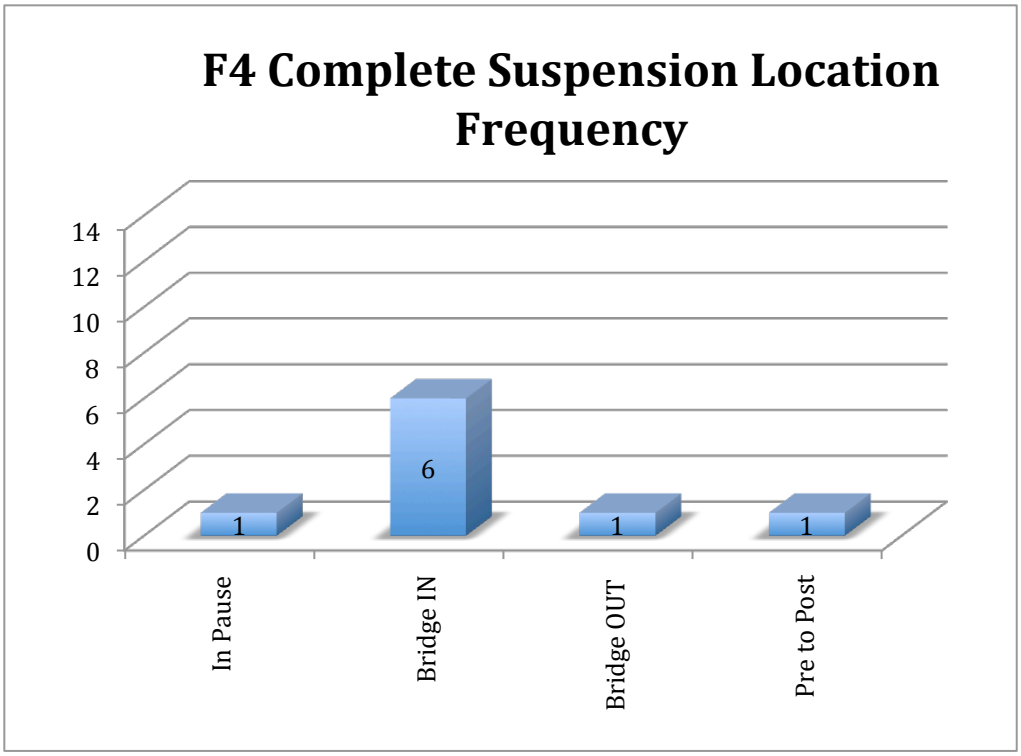


Figure 61. F4 complete suspension location frequency

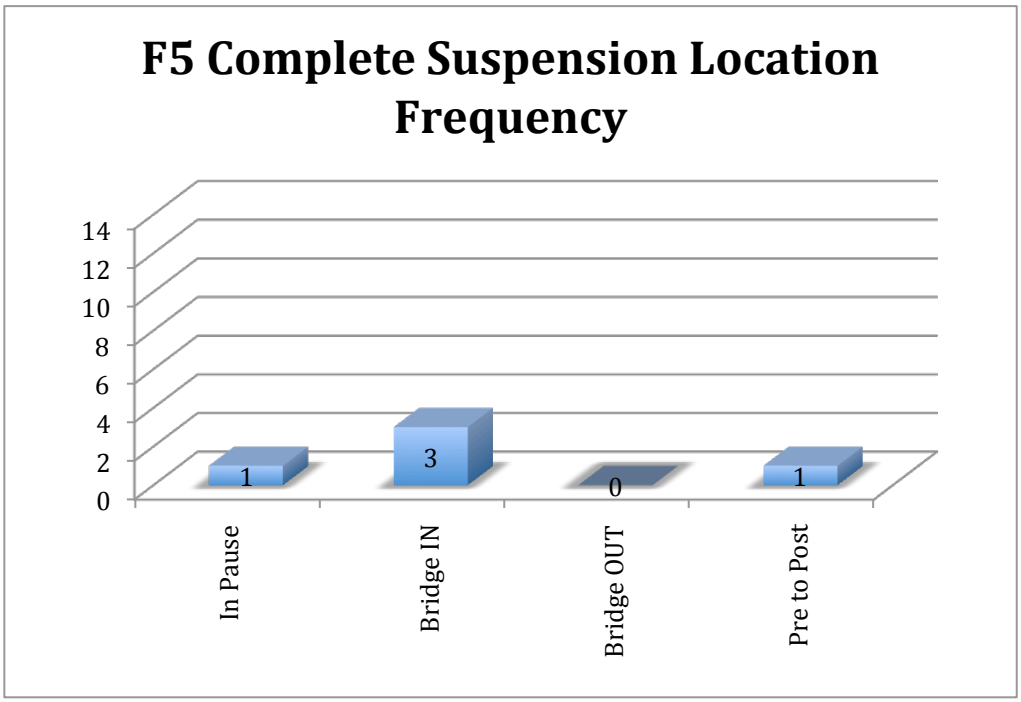


Figure 62. F5 complete suspension location frequency

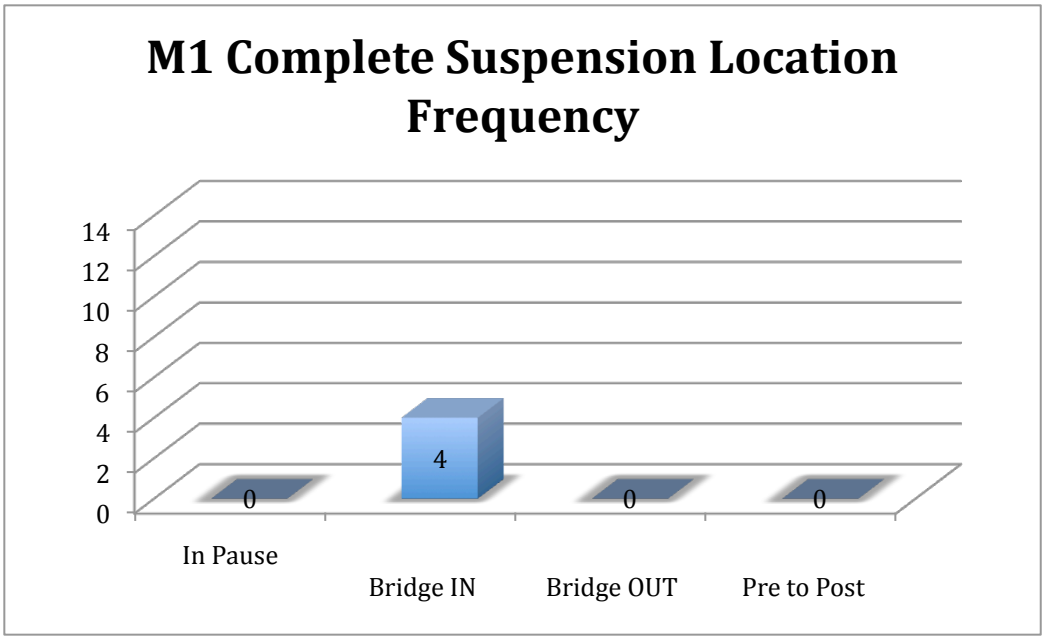


Figure 63. M1 complete suspension location frequency

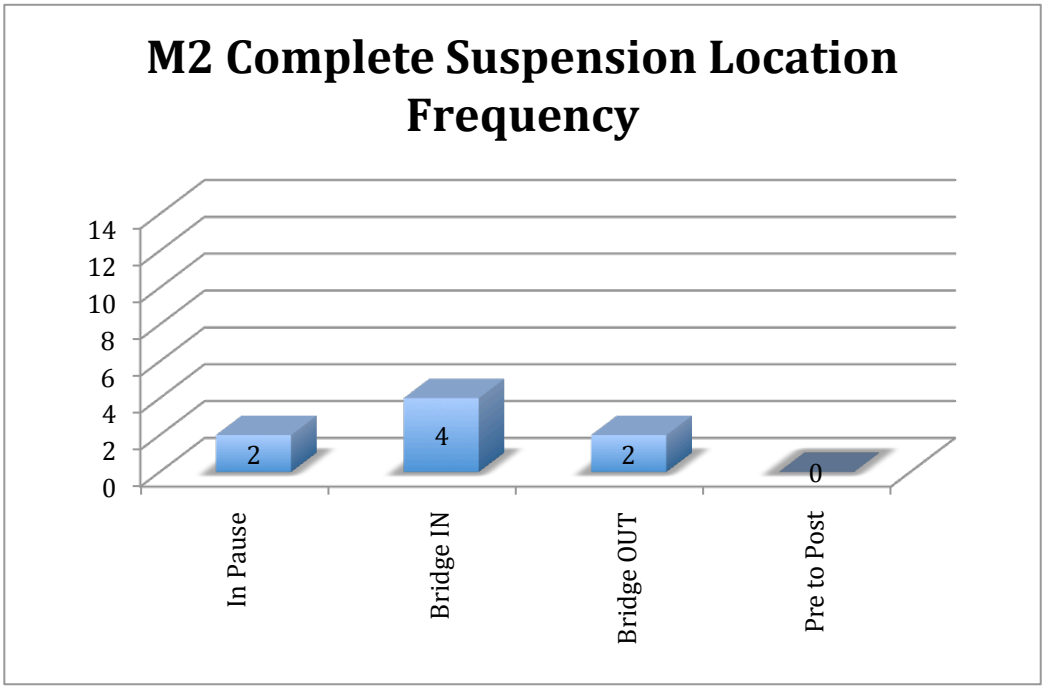


Figure 64. M2 complete suspension location frequency

Results indicate that for all subjects but F2, gesture suspensions in pause regions occur more often bridging into pauses than in any other location. Combined with results from section 4.5, this implies that speakers suspend gesturing before a pause and into a

pause, but may begin gesturing again as they bridge out of the pause. Subject F2's suspensions were distributed fairly evenly across the bridging in, bridging out, and pre-to-post locations.

4.4.4 Complete phase locations: behavior at pause areas

Complete gesture phase (preparation, stroke, release) locations were also calculated, in order to determine the distribution of gesture phases in comparison with the distribution of regular gestures.

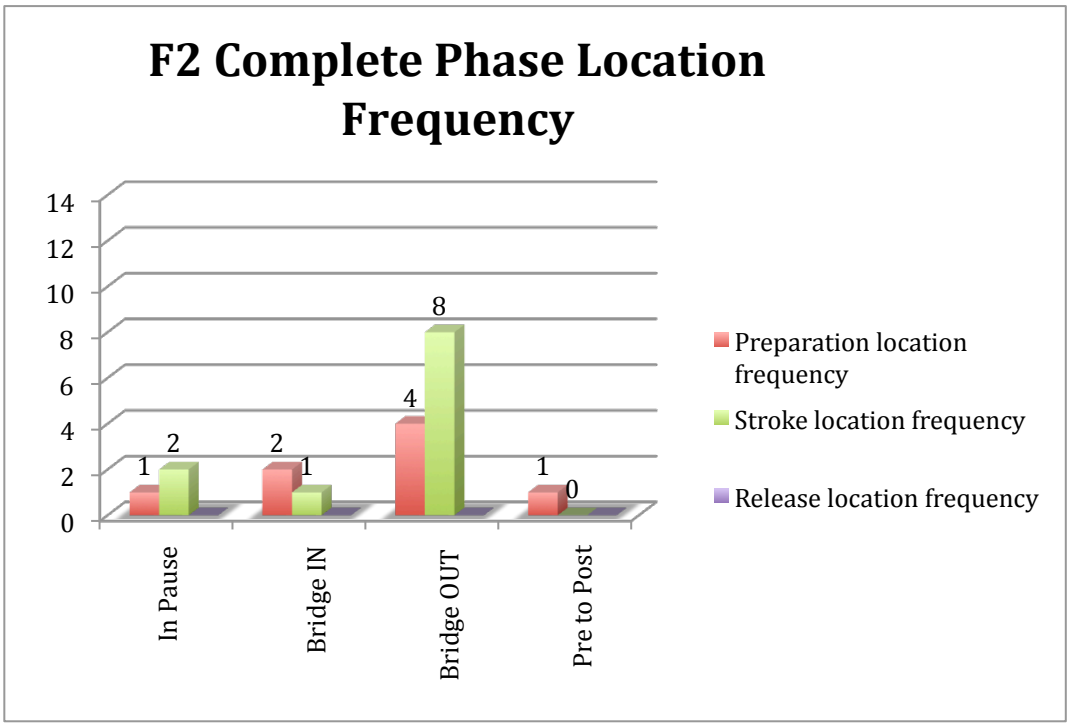


Figure 65. F2 complete phase location frequency

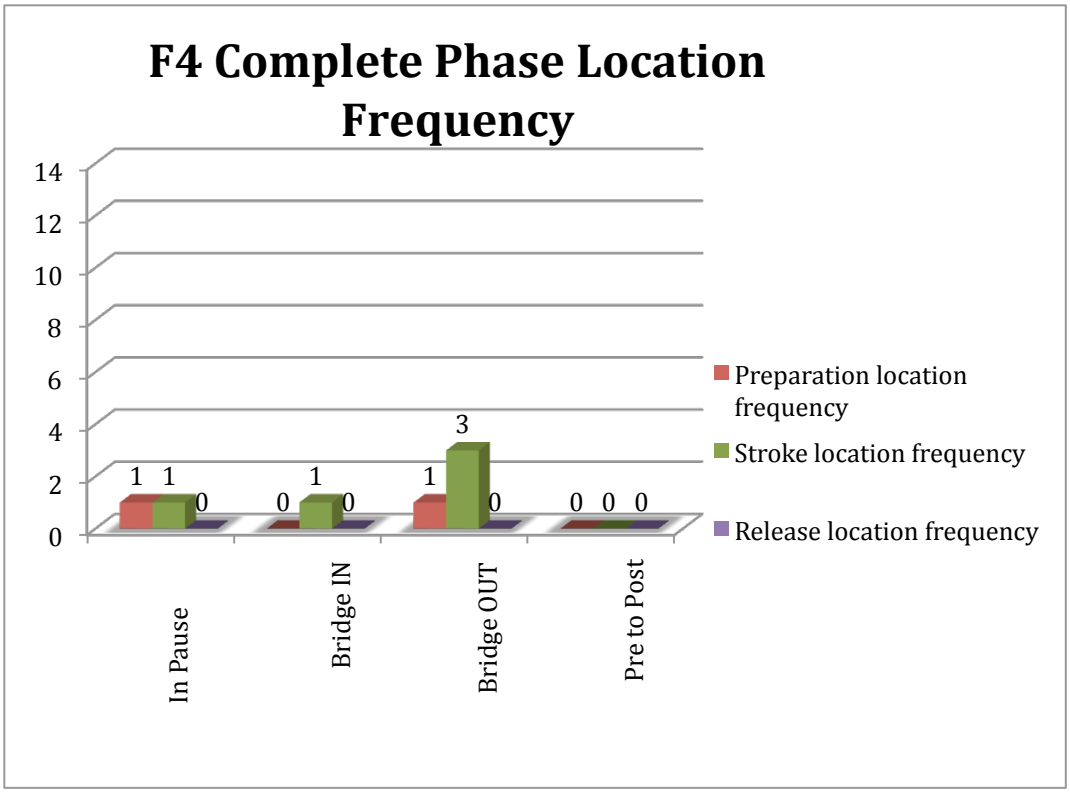


Figure 66. F4 complete phase location frequency

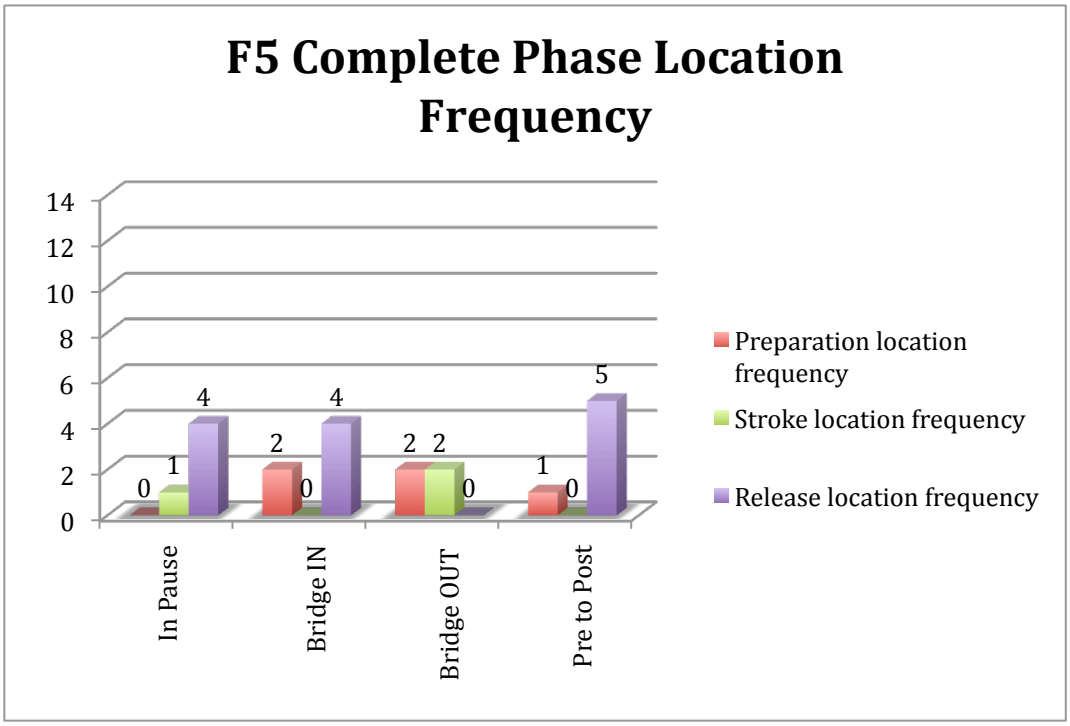


Figure 67. F5 complete phase location frequency

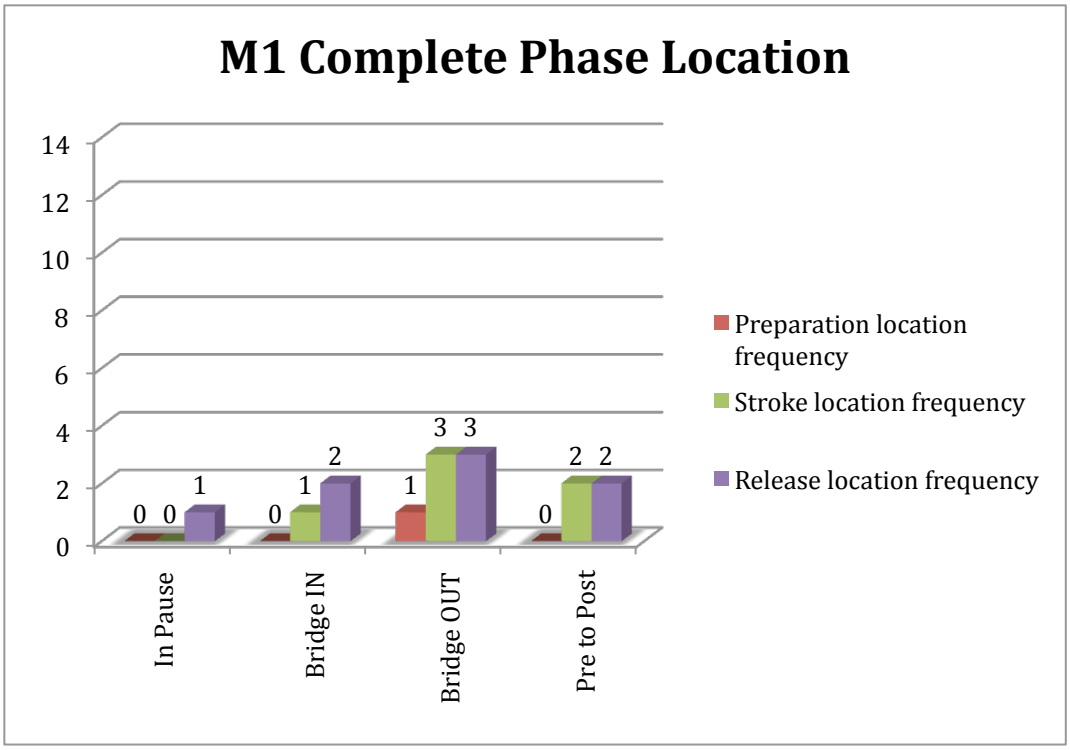


Figure 68. M1 complete phase location frequency

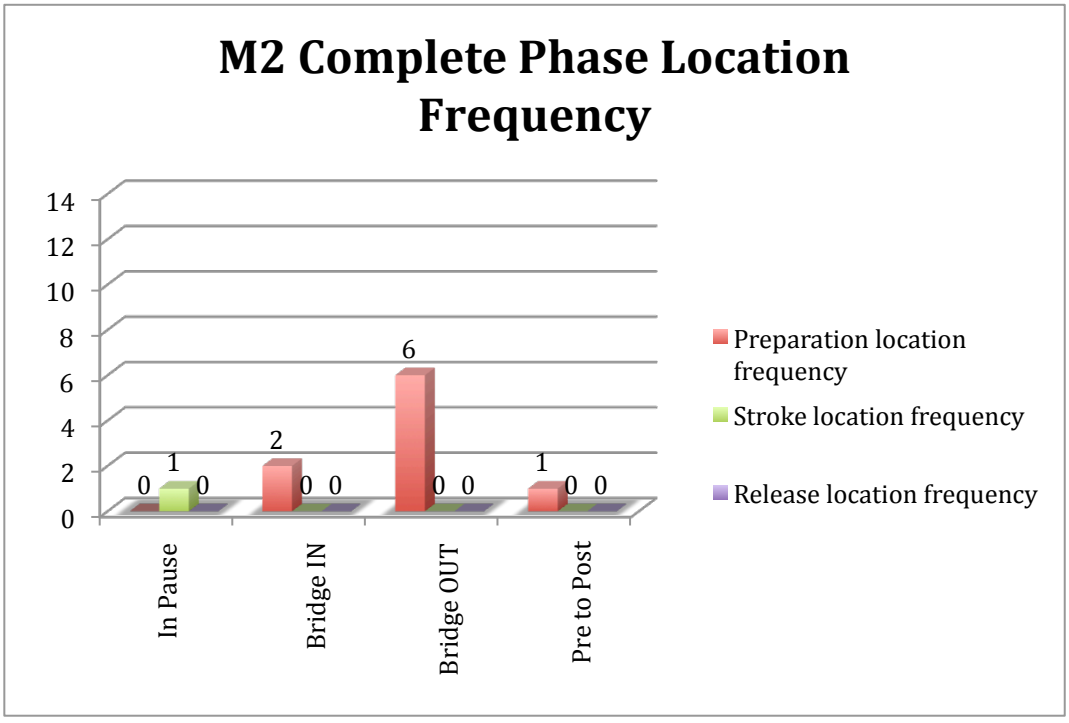


Figure 69. M2 complete phase location frequency

Results indicate that gesture phase locations at pauses differ across individual speakers. Subject F2 showed a high frequency of strokes bridging out of the pause, and had no release phases labeled around pauses. Subject F4 showed a similar pattern, but with fewer phases labeled overall. F5 showed an approximately equal frequency of releases across pauses, bridging in to the pause, and pre-to-post pause, but no releases occurred bridging out of the pause. Four out of the five gestures occurring in F5's pauses were labeled as release phases. F5's preparation and stroke phases were distributed across the pause area locations. Subject M1 showed a slightly greater frequency of strokes and releases bridging out of the pause than in other locations. M1 only showed one preparation phase labeled in pause areas, occurring bridging out of the pause. M1's only gesture occurring within the pause was labeled as a release phase.

In conclusion, gesture phase locations differed across subjects. This could be due to individual variations across speakers, or it could also be due to the subjective nature of labeling gesture phases, as discussed in Section 2.2.3.

4.5 Gesture differences in disfluent vs. fluent pauses

Finally, gesture onset and target behaviors as well as gesture suspensions were calculated in pauses labeled by the naïve listener as fluent or disfluent (see Section 3.2.4). Subject F3 was once again excluded from the following data reports due to extremely low frequency of onsets and targets in pauses (see Tables 2 and 3).

4.5.1 Disfluent vs. fluent region durations

For each subject, the following durations are listed in Table 4 below: total speech excerpt duration, total pause duration (the sum of all pause durations within the speech excerpt), total disfluent pause duration (the sum of all pause durations listener marked as disfluent), total fluent pause duration (the sum of all pause durations listener marked as fluent), pre-disfluent pause duration, pre-fluent pause duration, post-fluent pause duration, and post-fluent pause duration.

Subject	Speech Excerpt	Total Pause Duration	Total Disfluent Pause	Total Fluent Pause	Total Pre-D	Total Pre-F	Total Post-D	Total Post-F
F2	68.318	10.349	6.7	3.649	5.459	5.233	3.977	4.324
F4	78.678	23.235	19.816	2.419	11.317	1.791	7.172	0.859
F5	79.896	16.572	11.53	6.965	8.854	4.582	5.918	1.797
M1	52.266	7.984	5.065	2.919	3.62	3.374	4.054	1.963
M2	145.078	33.995	22.311	11.694	8.253	7.973	6.063	4.136

Table 5. Speech region durations, divided into disfluent (D) and fluent (F) categories

All pauses were divided unequally into disfluent and fluent pauses, with disfluent pauses perceived by the listener much more frequently than fluent pauses. This is intuitive given the speech is spontaneous and speech interruptions occur often. For some subjects, the distribution of pause categorization is more disproportionate than in others (see for example F4's distribution in Table 5).

4.5.2 Raw gesture data for disfluent vs. fluent regions

For each subject, the following raw gesture data is listed below: in Table 5, onset frequency occurring in disfluent pauses, fluent pauses, pre-disfluent regions, pre-fluent regions, post-disfluent regions, and post-fluent regions; and in Table 6, target frequency occurring in the same speech regions respectively.

Subject	Onset Frequency					
	In D Pause	In F Pause	Pre-D	Pre-F	Post-D	Post-F
F2	6	5	4	3	7	4
F4	15	2	2	1	7	2
F5	8	4	7	5	6	3
M1	7	7	5	5	4	4
M2	12	4	5	5	4	4

Table 6. Onset frequency in speech regions, divided into disfluent (D) and fluent (F) categories

Subject	Target Frequency					
	In D Pause	In F Pause	Pre-D	Pre-F	Post-D	Post-F
F2	4	5	6	6	8	11
F4	7	1	1	2	9	2
F5	7	5	7	3	8	4
M1	4	3	3	3	9	5
M2	7	2	7	8	10	7

Table 7. Target frequency in speech regions, divided into disfluent (D) and fluent (F) categories

4.5.3 Gesture rate: disfluent vs. fluent regions

Gesture onset and target rates were calculated for disfluent pause regions and fluent pause regions. Results are given by subject in Figures 70-74 below.

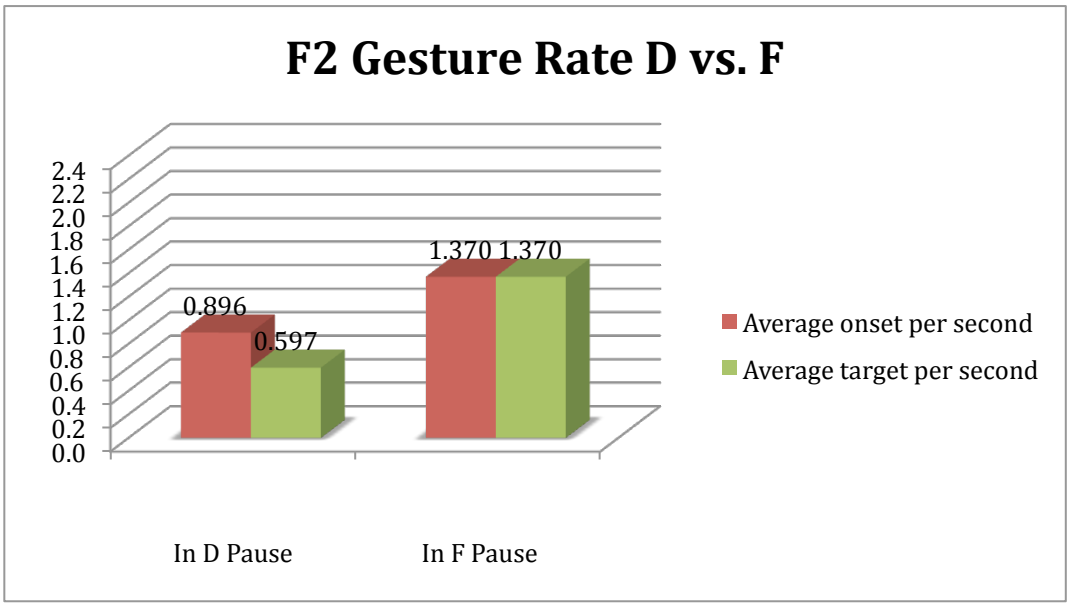


Figure 70. F2 average gesture onset and target per second

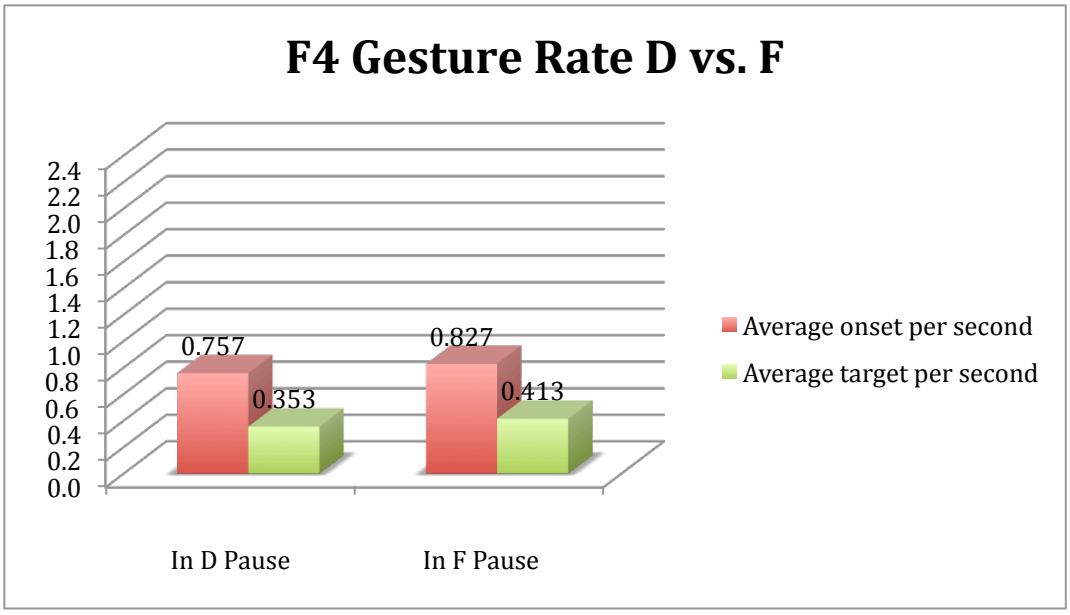


Figure 71. F4 average gesture onset and target per second

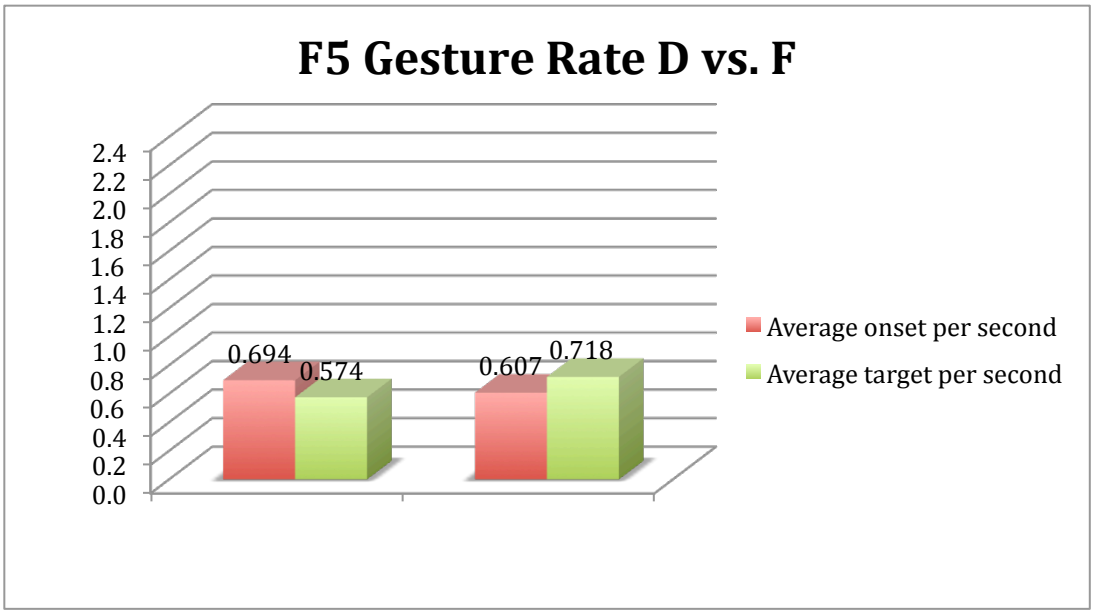


Figure 72. F5 average gesture onset and target per second

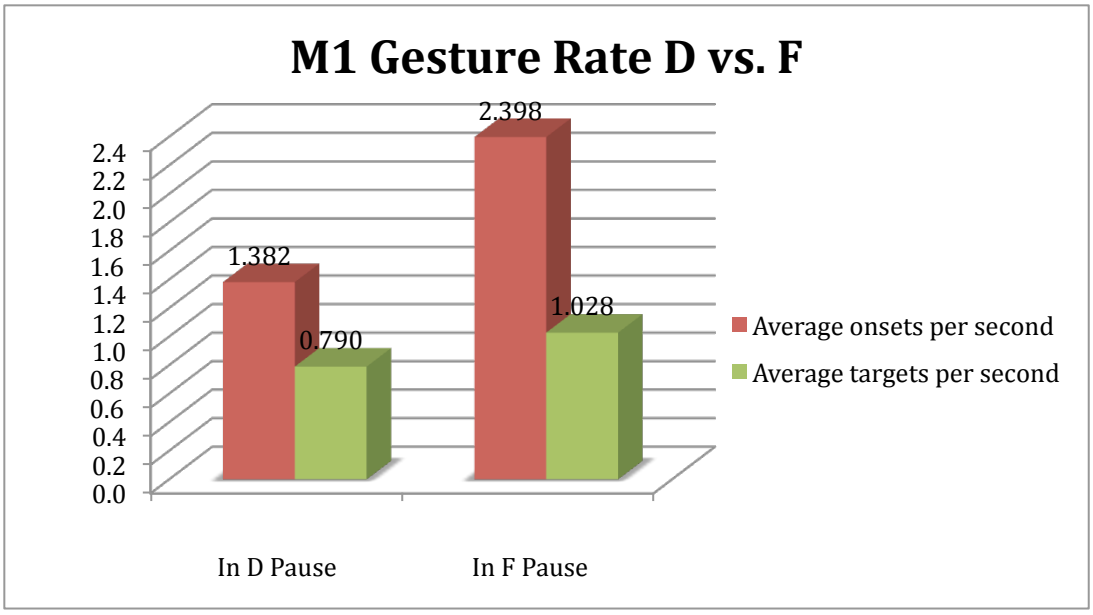


Figure 73. M1 average gesture onset and target per second

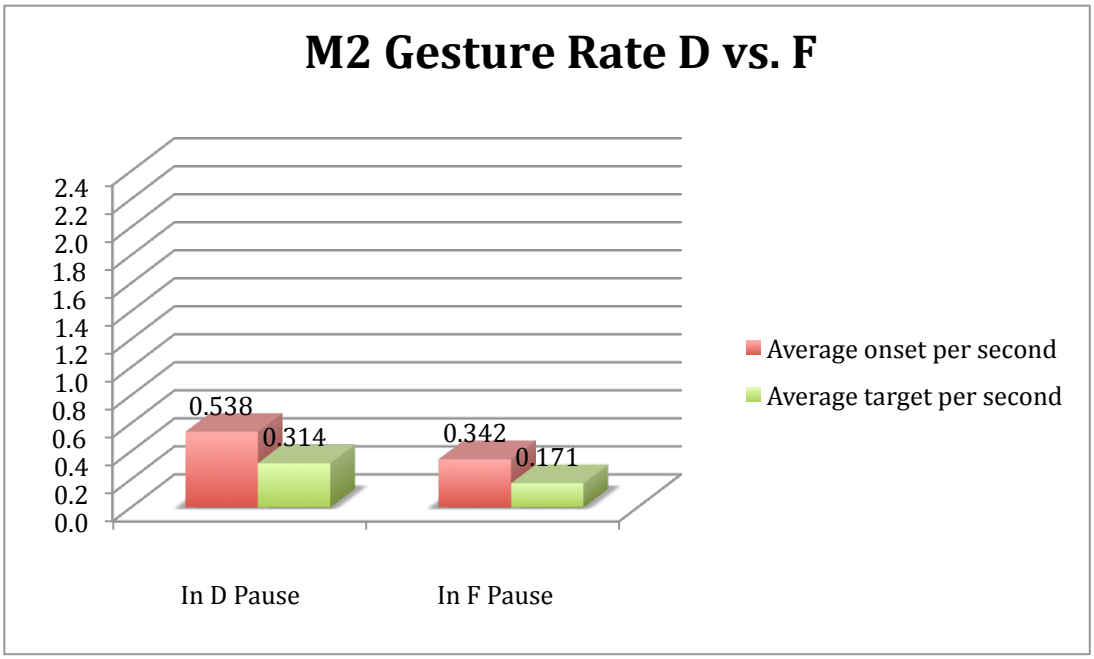


Figure 74. M2 average gesture onset and target per second

Results indicate that in some subjects (F2, F4, M1), gesture onsets and targets occur at a lower rate in pauses perceived as disfluent compared to pauses perceived as fluent. However, in other subjects (F5, M2), onsets and targets occur at an approximately equal rate or at a higher rate in pauses perceived as disfluent compared to pauses

perceived as fluent. The low number of fluent pauses labeled across subjects (as discussed in Section 4.8.1) makes it difficult to discern actual patterns in disfluent versus fluent pauses. However, according to these results, there does not appear to be a gestural behavior difference between the two pause types.

4.5.4 Suspension rate: disfluent vs. fluent regions

Average suspension rates were also calculated in order to determine if any pattern would emerge between suspension occurrence and disfluent versus fluent pauses. Suspension onset and target rates are shown by subject in Figures 75-79.

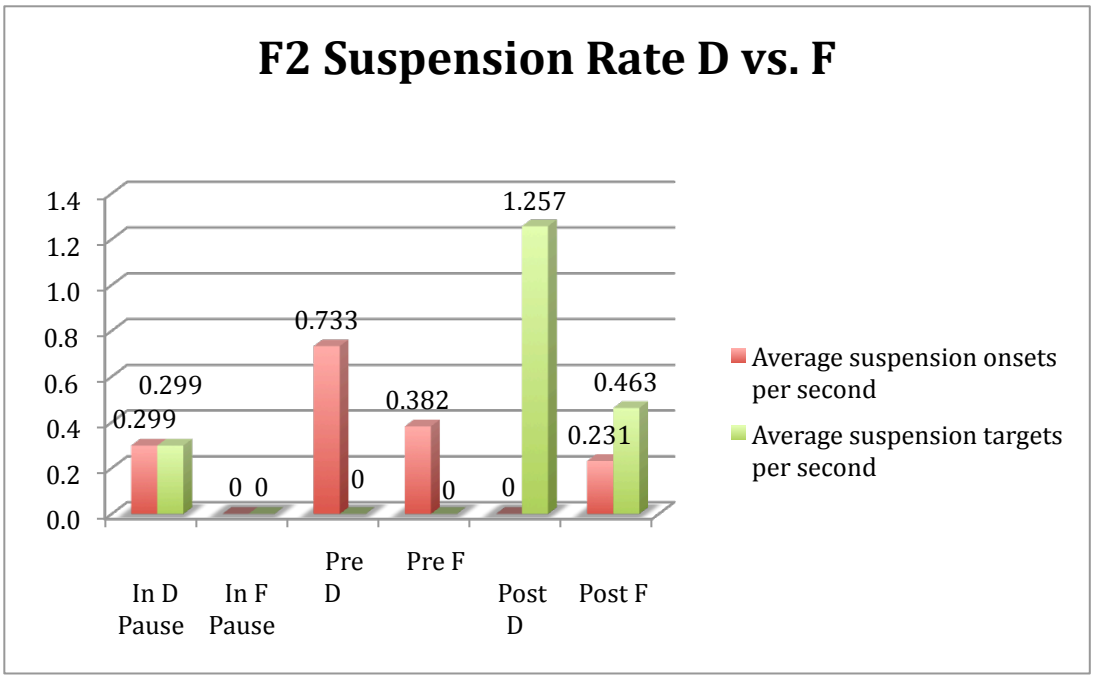


Figure 75. F2 average suspension onset and target per second

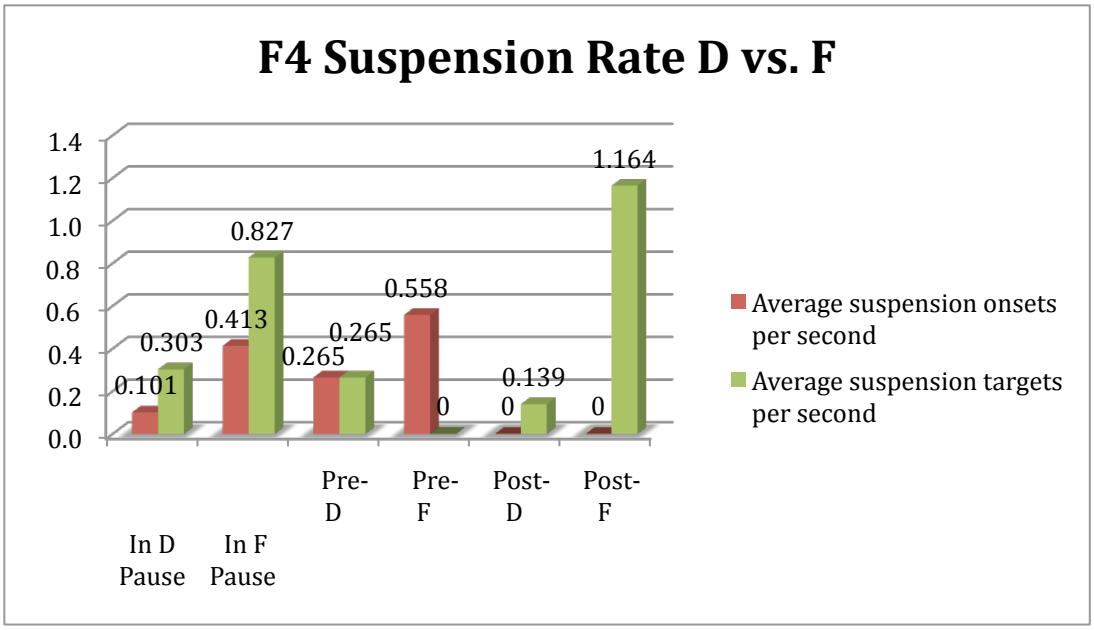


Figure 76. F4 average suspension onset and target per second

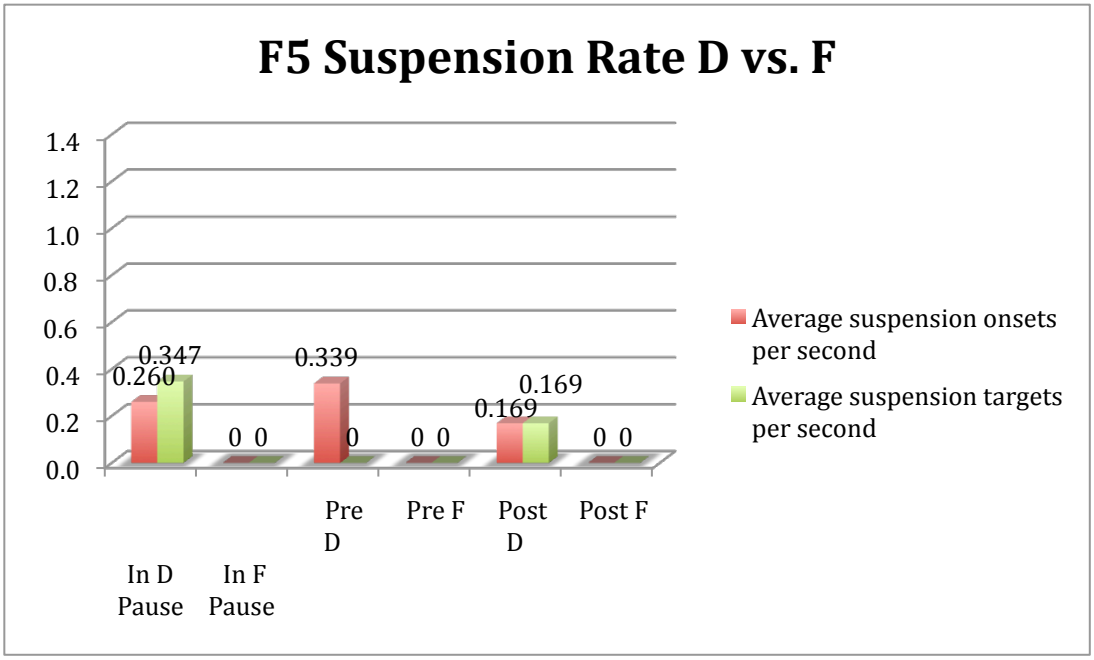


Figure 77. F5 average suspension onset and target per second

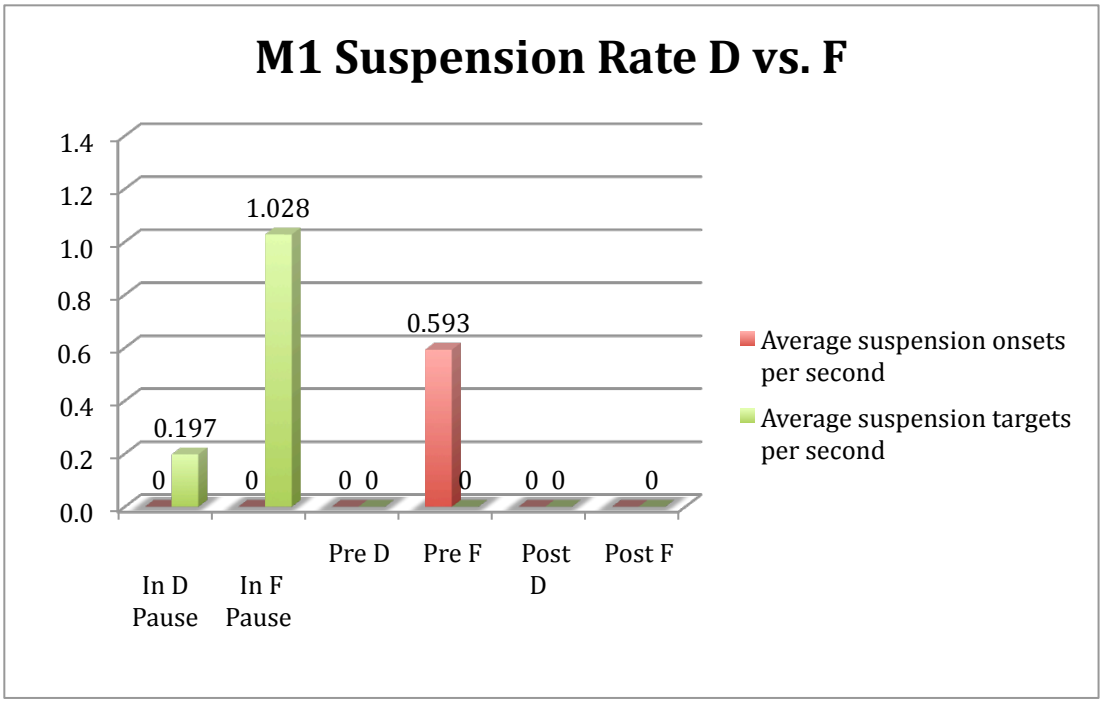


Figure 78. M1 average suspension onset and target per second

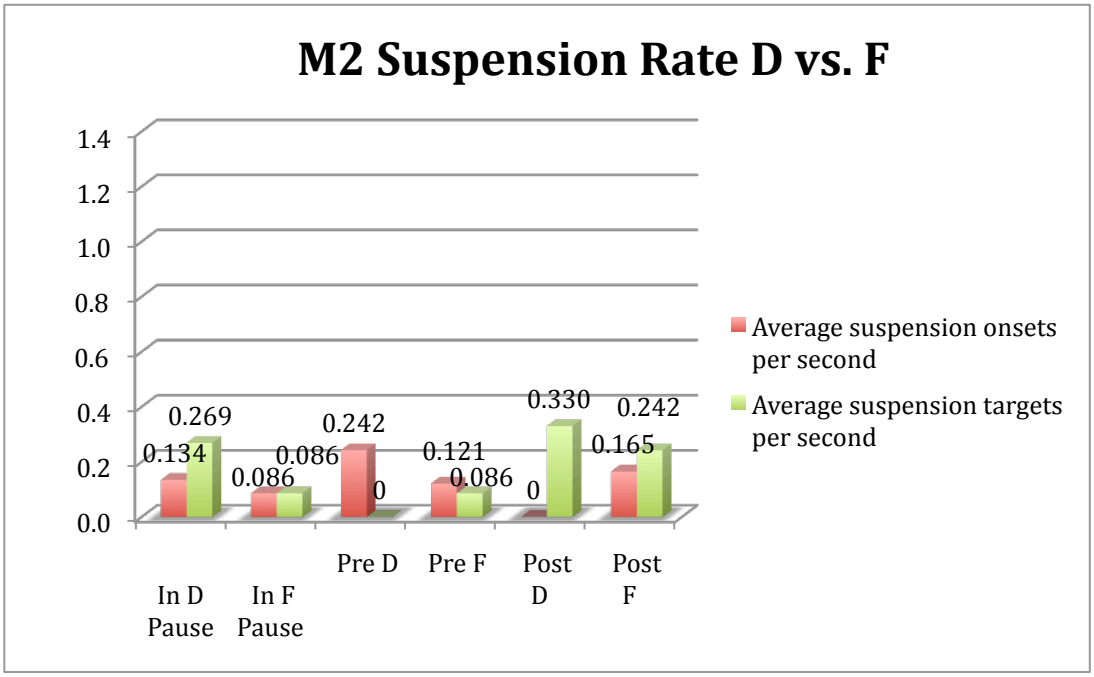


Figure 79. M2 average suspension onset and target per second

Results do not indicate a clear difference in suspension onset and target between pauses perceived as disfluent and pauses perceived as fluent. Some subjects (F2, F5, M2)

showed slightly greater suspension onset and target rates in perceived-disfluent pauses than in perceived-fluent pauses. Other subjects (F4, M1) showed the opposite pattern. F4 and M1 also possess the lowest total duration of pauses labeled as fluent, which could affect the results.

5. Discussion

The results shown in Section 4 indicate numerous conclusions about gesture behavior at pauses compared to fluent speech. Sections 4.3.1 and 4.3.3 found that gesture onsets and targets appear to occur slightly less frequently in pauses than in fluent speech, which implies that speakers begin and end gestures less frequently in pauses. However, the prevalence of onsets and targets in pauses indicates that there is significant activity in pause areas, even if it is slightly less than in fluent speech.

Sections 4.3.2 and 4.3.4 found that gesture onsets and targets appear to occur at a slightly lower rate in the pre-pause region than in fluent speech. This implies that speakers are less likely to begin or end gesturing just before a pause than in fluent speech with no upcoming pause. Conversely, in the post-pause region, onsets and targets appear to occur at a slightly higher rate than in any other speech regions. This implies that speakers are more likely to begin or end gesturing immediately following a pause than anywhere else in speech. Targets occurring at a higher frequency in the post-pause region indicate that gestures do begin either in the preceding pause or in preceding fluent speech bridging across the pause.

Section 4.3.5 examined gesture suspension onset and target rate across the four speech regions. It found that gesture suspension onsets and targets were more prevalent in pauses and pre- and post-pause regions, although they did also occur in fluent speech for some subjects, especially the two hand gesturers, F4 and M2. Onsets consistently occurred more in the pre-pause region, while targets were distributed differently for different subjects. This result indicates that suspensions often start just before a pause and may bridge into the pause, across the pause into the post-pause region, or into following speech.

Section 4.4 examined complete onset-target gesture units. Results from section 4.4.1 indicate that despite a prevalence of onsets and targets in pauses, complete gesture

units occur far less frequently in pauses than in fluent speech. This implies that gestures more often bridge in or out of the pause than occur completely within it. Section 4.4.2 examined complete gesture behavior at pause areas, and found that complete gestures at pauses occur most commonly bridging out of the pause. Section 4.4.3 found that complete gesture suspensions occur most commonly bridging into the pause. Section 4.4.4 did not find significant results for gesture phase location in relation to the pause.

One reason for the bridging-out-of-pause behavior could be that gestures may serve as a cognitive bridge between pauses and fluent speech, helping the speaker to flow seamlessly back into fluent speech. Gesture suspensions occur more commonly going into the pause, and may thus serve as gestural indications of a disruption in the speech flow.

Another reason for the bridging-out-of-pause behavior could be due to the fact that gestures appear to be rhythmically based and linked with the prosodic structure of speech; for instance, gesture target and pitch accent have been shown to be aligned across several languages (see Yassinik et al 2004, Loehr 2004). Therefore, when linguistic structure stops, i.e. in pauses, gestures also stop, until the speech starts again. If upcoming speech contains a pitch accent that just follows a pause, gesture onset may begin in the pause in order to allow for target-pitch accent alignment. This would indicate gestural planning in parallel with linguistic planning. This could also explain the result that suspensions occur commonly going into the pause, as the gesture suspends to wait for its speech counterpart to get back on track. One final reason for this bridging behavior could be that gesture onset may in fact align with articulatory onset, as opposed to acoustic onset. This is due to the fact that articulatory movements begin before acoustic onset. However, further studies would need to be done both in speech articulation and gesture onset time to determine this.

Section 4.5 examined whether gestural differences existed between disfluent and fluent pauses. No clear pattern emerged for either regular gestures or gesture suspensions. This could be for several reasons. First, the low number of fluent pauses labeled across subjects in comparison with the higher number of disfluent pauses makes it difficult to discern patterns in disfluent versus fluent pauses. Second, the perception study for disfluent versus fluent pauses was subjective and served merely as a guideline

for probable pause categories. Pauses however may belong to the other category, and as such results may be misleading. Further studies would need to be done with a larger sample so that enough pauses perceived as fluent could be gathered for examination. The hypothesis for a study with a larger sample size is that in disfluent pauses, gesture suspensions will be prevalent, followed by a resumption of gesturing bridging out of the pause, whereas in fluent pauses, targets of previous gestures, especially targets of release phases, will occur more prominently. This is due to observation of many speakers who demonstrate an apparent connection between abrupt, perceived-as-disfluent pauses and suspensions, especially freezes; as well as a connection between fluid, natural pauses and gesture releases.

Figure 80 illustrates gesture behavior at the pause versus in the corresponding fluent area. Combined results from all of the above sections indicate that while some complete gestures do occur in pauses (shown by the semi-transparent arrow in the pause region), most often gestures bridge into or out of the pause, most notably bridging out (shown by slightly bolder arrows). Additionally, gesture suspensions (shown by the green arrow) occur most often bridging into the pause. Conversely, in fluent speech, gestures appear to occur equally as often throughout the entire region. Because complete gestures do not occur often in pauses, a tentative conclusion would posit that gestures are planned in accordance with linguistic structure and parallel the structural properties of speech.

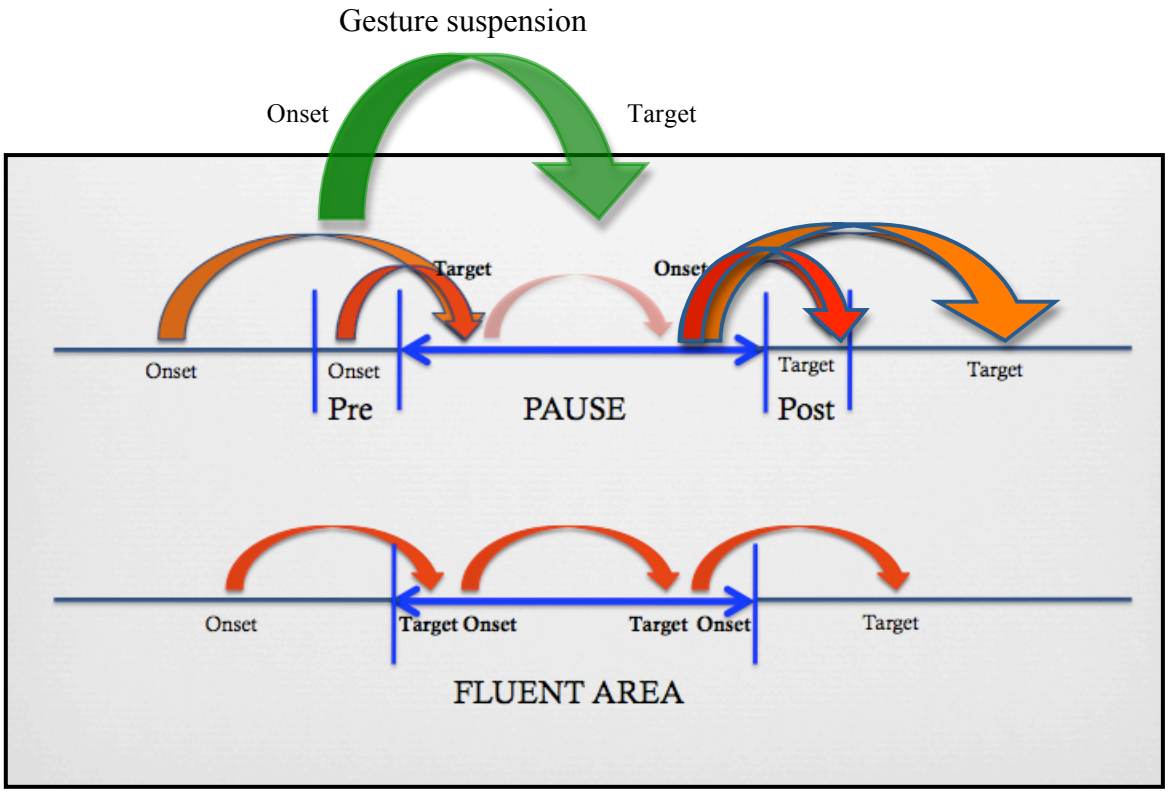


Figure 80. Visual representation of gesture behavior at the pause versus in fluent speech. At a pause, gestures often occur bridging into or out of the pause, more commonly out of the pause (shown by bolder arrows leading out of the pause). Gesture suspensions occur most commonly bridging into the pause (shown by the green arrow). Conversely, in fluent speech, gestures are equally likely to occur in all locations.

6. Enhanced labeling method for future studies

An additional goal of this study is to improve the reliability of gesture labeling for future studies. The following method was developed as an improvement of the labeling system used in this study; it has not yet been used extensively. The method was applied to a new subject, M3, who was not included in this study's data set. This method moves away from subjective labeling and towards semi-automatic labeling, based on gesture velocity information.

M3's video was imported into After Affects, a video effects software, and set up for motion tracking. Motion tracking involved selecting a high contrast region on the subject's face in the video, such as the eyebrow, and tracking the point through time. At each frame, or 1/60 of a second, x and y position data was recorded. This information was then exported out of After Effects into a text file. The text file, video file, and audio file of the subject's recording was then imported via a script (developed by Mark Tiede at Haskins Laboratories) into MView (M. Tiede under development) for Matlab, which

displays a visual representation of movement through time (see Figure 81). An MView plug-in automatically finds gestures based on this movement and notates them according to gesture onset and offset (see Figure 82a and b). These labels are based on gesture velocity information. For instance, when a gesture reaches zero velocity, this indicates it has either stopped or is changing direction, which indicates this is location of the target of previous movement and, if a changed direction, the onset of upcoming movement.

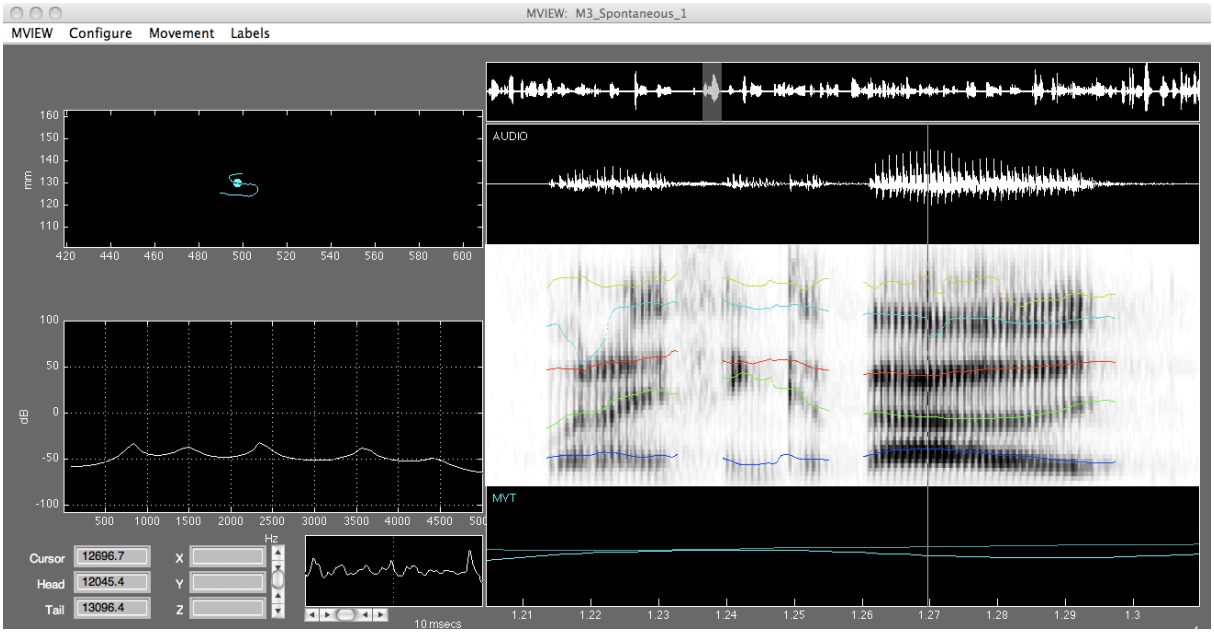
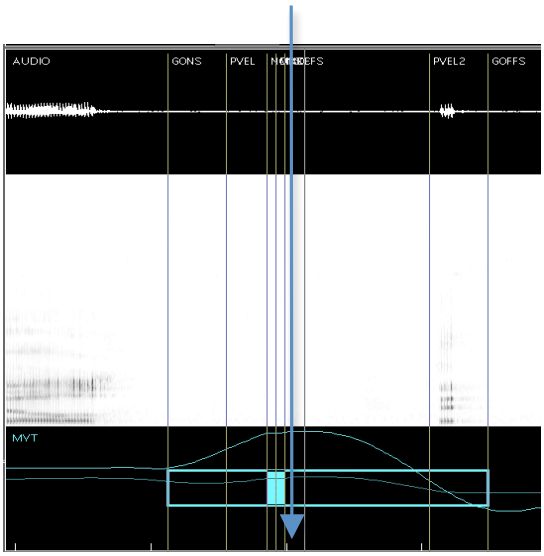


Figure 81. Screenshot of MView set-up for new subject. 1. Entire audio waveform of subject's recording, 2. Zoomed in excerpt of highlighted audio in 1, 3. Formants, 4. Movement through time, 5. x and y position location for tracker point at each frame (point moves position as cursor moves through excerpt) 6. Amplitude



a. Another audio excerpt with prevalent head movement

b. Video accompaniment of a. at arrowhead

Figure 82a and b. Subject's head in b. is at peak of curve in a., i.e. is at zero velocity; this indicates a gesture direction change

In conclusion, this new method involves tracking subject motion through time and inputting tracker position data to Mview, which detects gesture onset, peak, and offset based on velocity minima and maxima. This new method of labeling gesture behavior via motion tracking and velocity data serves as an inexpensive, accessible alternative to the more elaborate three-dimensional motion capture, and is a vast improvement from manual methods.

7. Conclusion

This study elicited spontaneous, monologue speech from six subjects and recorded hand and head gesturing. Results indicate that complete gesture onset-target units do not often occur in pauses, but more often gestures bridge into the pause from previous speech or out of the pause into following speech. Gestures at pause areas appear to occur most commonly bridging out of pauses. This could indicate that gestures serve as cognitive connectors between pauses and fluent speech. It could also indicate that in

pauses as well as in fluent speech, gestures are planned so as to align with certain structural properties in following speech, such as pitch accents.

Gesture suspensions occur in pause regions and in fluent speech, but most commonly occur bridging into the pause. This could indicate that gestures stop in anticipation of speech stopping, and that the two systems are thus structurally linked and planned together. However, this would require further examination.

Disfluent and fluent pauses were also examined for gestural differences but there was not enough conclusive data to make a claim about the relationship between gesture behavior and pause type.

Lastly, a new method of motion tracking video data and automatically labeling gesture was developed and is still being refined for use in future studies. This method, if proven successful on a larger amount of data, would contribute to the accuracy and efficiency of future gesture studies, especially those that cannot afford a three-dimensional motion capture system.

8. Special Thanks

I would like to thank all those amazing people who supported me throughout the development of my senior thesis. Thank you to my parents, who encouraged me to keep trucking even when it got overwhelming, and who support me to the end. Thank you to Sean, who helped me immeasurably with automating parts of the analysis process, and who kept me going with never-ending kindness (and a constant supply of snacks!).

Thank you also to my fellow linguistics majors who politely listened to me present my developing experiment in class at least fifteen times. Thank you to the Linguistics Department and all my linguistics professors who have enriched my experience at Yale. Thank you especially to Raffaella Zanuttini, who has been a kind and helpful guide throughout my navigation of the Linguistics major.

And thank you to my senior thesis advisor, Jelena Krivokapic, who I am forever indebted to for teaching me so much about scientific research, quality writing, and constant perseverance. I am grateful to have worked with you for the past two years, and without you I never could have accomplished even a fraction of what appears in this paper.

9. Bibliography

Butterworth, B. and Beattie, G., 1978. “Gesture and silence as indicators of planning in speech.” In R.N. Campbell & P.T. Smith, eds., *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*. New York: Olenum Press.

Cassell, Justine, David McNeill, and Karl-Erik McCullugh, 1999. "Speech-gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information." *Pragmatics and Cognition* 7.1: 1-34.

Cave, C., I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, 1996. "About the Relationship between Eyebrow Movements and F0 Variations." *Spoken Language* 4: 2175-178.

Cooper, William E., and Jeanne Paccia-Cooper, 1980. *Syntax and Speech*. Harvard College. Print.

Fernanda, Ferreira, 1991. "Effects of Length and Syntactic Complexity on Initiation times for Prepared Utterances." *Journal of Memory and Language* 30.2: 210-33.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., Bernstein, L., 2003. "Optical phonetics and visual perception of lexical and phrasal stress in English." *Proceedings of the ICPhS.* Barcelona, Spain: 2071-2074.

Kendon, Adam, 1980. "Gesticulation and Speech: Two Aspects of the Process of Utterance." *The Relationship of Verbal and Nonverbal Communication*. Comp. Mary Ritchie. Key. The Hague, Mouton: 207-227.

Kendon, Adam, 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. Print.

Krahmer, Emiel, and Marc Swerts, 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.

Loehr, Daniel P, 2004. "Gesture and Intonation." (Dissertation).

McClave, Evelyn, 1997 (published 1998). "Pitch and Manual Gesture." *Journal of Psycholinguistics Research* 1st ser. 27: 69-89.

McNeill, David, and Elena Levy, 1982. "Conceptual Representations in Language Activity and Gesture." *ERIC (Education Resources Information Center)*.

McNeill, David, ed., 2000. *Language and Gesture*. Cambridge University Press. Print.

McNeill, David, 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago. Print.

McNeill, David, 1985. "So You Think Gestures Are Nonverbal?" *Psychological Review* 92.3: 350-71.

Moscovici, Serge, 1967. "Communication Process and Language." Ed. L. Berkowitz. *Advances in Experimental Social Psychology* 3: 225-70.

Rochester, S. R., 1973. "The Significance of Pauses in Spontaneous Speech." *Journal of Psycholinguistics Research* 2: 51-81.

Sassenberg, Uta, and Elke Van Der Meer, 2010. "Do We Really Gesture More When It Is More Difficult?" *Cognitive Science* 34: 643-64.

Seyfeddinipur, Mandana, 2006. "Disfluency: Interrupting Speech and Gesture." (Dissertation).

Treffner, Paul, Mira Peter, and Mark Kleidon. "Gestures and Phases: The Dynamics of Speech-Hand Communication." *Ecological Psychology* 20.1 (2008): 32-64.

Tuite, Kevin, 1993. "The Production of Gesture." *Semiotica* 93: 83-105.

Werner, Heinz, and B. Kaplan, 1963. "Symbol Formation: An Organismic-developmental Approach to Language and the Expression of Thought."

Yasinnik, Yelena, Stefanie Shattuck-Hufnagel, and Nanette Veilleux, 2005. "Gesture Marking of Disfluencies in Spontaneous Speech." *Disfluency in Spontaneous Speech Workshop*: 173-78. *ISCA Archive*.

Yasinnik, Yelena, Margaret Renwick, and Stefanie Shattuck-Hufnagel, 2004. "The Timing of speech-accompanying gestures with respect to prosody." *From Sound to Sense*: 97-102.