TEMPORAL RELATIONS OF VERBAL AND NON-VERBAL BEHAVIOR IN STORYTELLING

by

MATTHEW STAVE

A DISSERTATION

Presented to the Department of Linguistics
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: Matthew Stave

Title: Temporal Relations of Verbal and Non-Verbal Behavior in Storytelling

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree by the Department of Linguistics by:

Eric Pederson	Chair
Vsevolod Kapatsinski	Core Member
Melissa Baese-Berk	Core Member
Spike Gildea	Core Member
Zhuo Jing-Schmidt	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2018

DISSERTATION ABSTRACT

Matthew Stave
Doctor of Philosophy
Department of Linguistics
September 2018

Title: Temporal Relations of Verbal and Non-Verbal Behaviors in Storytelling

This dissertation takes a ‘big data’ approach to analyzing a corpus of multimodal storytelling with the goal of providing data for researchers interested in developing more holistic models of production that integrate verbal and non-verbal behavior. Rather than approaching the data with a specific hypothesis in mind, I approach the data with a set of methods that analyze the temporal relationship between two behaviors and apply the methods to every single possible pair of behaviors. Rather than using the data to test hypotheses, I am using it to formulate them.

The methods used in this dissertation examine covariation between behaviors (how much do two behaviors overlap with each other, and is this more or less likely than we would expect, given a random distribution of the two behaviors), the sequential patterns of behaviors (the multimodal n-grams of behaviors that are most strongly associated), and the frequency distribution of behavior boundaries (the timing of behavior onsets and offsets near other behavior onsets and offsets).

The analyses examine all possible pairs of behaviors from four modalities (head gesture, manual gesture, eye-gaze, and speech), as well as looking within and across roles of speaker and listener. A list of testable hypotheses is given, based on the findings in the data.

CURRICULUM VITAE

NAME OF AUTHOR: Matthew Stave

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon

George Fox University, Newberg, Oregon

DEGREES AWARDED:

Doctor of Philosophy, Linguistics, 2018, University of Oregon

Bachelor of Arts, English / Spanish, 2001, George Fox University

AREAS OF SPECIAL INTEREST:

Psycholinguistics, gesture studies, multimodal communication, corpus linguistics, computational linguistics, and semantics

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, American English Institute, 2011-2014, University of Oregon, Eugene

Graduate Research Fellow, Department of Linguistics, 2014-2015, University of Oregon, Eugene

Graduate Teaching Fellow, Department of Linguistics, 2015-2018, University of Oregon, Eugene

GRANTS, AWARDS, AND HONORS:

Graduate Teaching Fellowship, University of Oregon, 2011-2018

Travel Grants, University of Oregon, 2011-2018

Institute of Cognitive and Decision Sciences Dissertation Research Award, University of Oregon, 2018

PUBLICATIONS:

Stave, M., Pederson, E. (2018). Testing the psychology of a typological pattern: The nature of the IN/ON cline. Manuscript submitted for publication (*Linguistic Typology*).

Stave, M., A. Smolek, & V. Kapatsinski. (2013). Inductive bias against stem changes as perseveration: Experimental evidence for an articulatory approach to output-output faithfulness. Proceedings of the 35th Annual Meeting of the Cognitive Science Society, 3454-59. Austin, TX: The Cognitive Science Society.

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the support of my advisor, Eric Pederson, who has shown me wisdom and intelligence, and also honesty and kindness. He is a remarkable human being and I wish him happiness, health, and excellent grad students. I would also like to thank Volya Kapatsinski for being a reliable and brilliant mentor when I first found an interest in linguistics. Thanks also to my committee members, Melissa Baese-Berk, Zhuo Jing-Schmidt, Spike Gildea, and all the faculty at the UO Linguistics Department.

I would like to thank the friends who have made this PhD worth pursuing, including every grad student in the Linguistics Department, which is almost glutted with excellent people. I'm especially grateful to Shahar Shirtz for being an unwilling source of wisdom and compassion, Jaime Peña for many hours of distraction and pisco, Allison Taylor-Adams and Aaron Adams for insightful and uplifting conversation, Wan Vajrabhaya for being an excellent storyteller, Amos Teo for reliably barging into my office when I most needed it, Amy Smolek and Zara Harmon for examples of courage and sarcasm, respectively, Charlie Farrington for consistently excellent music recommendations, and Paul Olejarczuk, Manuel Otero, Tyler Kendall, and Charlotte Vaughn for the much needed hours of musical revelry.

I would like to thank Mogwai, Explosions in the Sky, Portishead, This Will Destroy You, Grimes, This Patch of Sky, God is an Astronaut, Hammock, and The Mars Volta for making music that distracted the parts of my brain that would otherwise have made writing impossible, and Tom Waits, Elliott Smith, Sufjan Stevens, Jason Isbell, Marvin Gaye, and Townes Van Zandt for distracting the rest of it.

Finally, I would like to thank my family. They are the most wonderful family.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Chapter III.

Chapter IV.

Figure Page

Figure Page

Figure Page

LIST OF TABLES

Table Page

Table Page

Table Page

Table Page

CHAPTER I: INTRODUCTION AND REVIEW OF THE LITERATURE

1. Introduction: The Coordination of Verbal and Non-verbal Behavior

Human communication is complex. Linguistic systems, one important component of human communication, allow for infinitely generative sequences of units that convey unique meanings. The units of language – phonemes, morphemes, words, and phrases, to name a few – combine in systematic, conventional ways to create a shared conduit for transmitting thought. We know as much as we do about the complexity of language in a large part because we have orthographic systems that allow us to document, share, and compare linguistic data, and we have orthographic systems in a large part because much of language can be decomposed into discrete, categorical units. It is not an accident that the continuous parts of language only began to receive attention after the categorical parts. Early systematic studies of language examined the kinds of things that texts made easy: historical change of structural units of morphology, syntax, and sound. It wasn't until audio recordings became affordable that the continuous characteristics of phonetics and prosody began to receive attention, and it wasn't until corpora became large and corpus searches became straightforward that the probabilistic properties of language began to be studied in earnest (aside from a few visionaries, such as Zipf (1949)). The continuous and probabilistic properties of semantics and pragmatics (e.g. the strength of associations between complex forms and conceptual meanings, or the degree of inappropriateness when a particular social convention is flouted with language) are still

mostly only available through introspective methods. Technology may follow science, but science also follows technology[1].

Over the past forty years, we have had increasingly affordable access to video technology. This has been invaluable for researchers of signed language systems, allowing the rapid recording and sharing of linguistic data. But it has also opened up the way for researchers to better explore another important set of pieces of human communicative behavior: non-verbal communication. Non-verbal communication involves all the behaviors produced during communication that are not part of the traditional linguistic system, including behaviors in the head, face, hands, eyes, and posture. Two characteristics that many non-verbal behaviors share is that they co-occur with (or in some way temporally coordinate with) speech, and that they can contribute to the semantics or pragmatics of speech in a variety of ways, such as by adding emphasis (a nod), focusing attention (a raised eyebrow), representing semantics (a tracing gesture), or offering a turn (an outstretched open palm) (McNeill 2000, Goldin-Meadow 1999). Another important characteristic is that non-verbal communication is universal – there are no language-users who do not employ these behaviors in natural language usage, even among the blind.

All of this leads to a prediction: researchers in the near future will be building holistic models of human communication that integrate a wide variety of aspects of both verbal and non-verbal behavior. Effectively, existing models of language production will be incorporated into larger models that explain both verbal and non-verbal behavior. This

[1] Credit to Eric Pederson.

seems probable for several reasons that have been discussed above. Non-verbal behavior is a universal characteristic of human communication, and must be considered by anyone interested in a complete description of communicative behavior. The close temporal integration of verbal and non-verbal behavior (McNeill 2000; more discussion in Section 1.6) suggests that both originate from a common source. And the technological barrier to describing visual behavior has grown much weaker, particularly with the increase of motion-capture systems. Incorporating non-verbal behavior into the field of linguistics would a natural next step, as the history of the field of linguistics has been one of adding more and more factors to explain how humans communicate with language (cognition, social identity, prosody, etc.) This dissertation is designed to help future researchers identify areas where the verbal and non-verbal systems interact, with the goal of facilitating the construction of holistic models of human communication. Specifically, I am looking at the temporal relations between verbal and non-verbal behaviors. The timing of production planning is foundational for understanding the nature of a communicative system, whether looking at the temporal co-variation of different aspects of the system (whether looking at the covariation of prosodic information with syntactic information, or looking at the covariation of non-verbal and verbal information).

I am also looking at these behaviors through the lens of multimodal communication, a discipline that looks at how communication is produced and comprehended in multiple modalities. These includes language, gestures, facial expressions, posture, and other means of expression. The term *modality*, as it is used in this dissertation, refers to any channel of expression that is used during communication – specifically, we will look at

the modalities of head gesture, eye-gaze, manual gesture, and speech, each of which interact with other modalities during communication, both within and across roles.

To illustrate how some of these behaviors interact during communication, let us take a look at four and a half seconds of story-telling narration, selected at random, in which a speaker (A) is describing a personal story to a listener (B), in which she was almost dashed against some jagged rocks while snorkeling in Hawai'i, before being saved by her father. The time course of each of the four modalities is shown below for both participants, and the brackets show the onsets and offsets of behaviors in each modality.

Example 1.

A Speech: [*I like kicked off the rock a little bit*] [*but*] [*luckily my dad came up and like*]
A Head: [small nod] [retract] [jut]
A Hands: [left hand lifted and thrown outwards]
A Gaze: ..gaze away][gaze towards][gaze away..

B Speech: [*mhm*]
B Head: [multiple nods]
B Hands:
B Gaze: ..gaze towards][gaze away][gaze towards..

There are four modalities shown here, and each modality interacts temporally with other modalities within the role of speaker/listener and across roles. What is more, they interact in predictable ways.

First we'll look at A. During her first speech segment, a very dramatic and spatiotemporally vivid event, she uses each of the three other modalities, each of them displaying temporal coordination with one or more other behaviors. She gives a small nod (down, then back up), and the downbeat of this nod occurs precisely at the vowel onset of the word *off*. At the same time, she is lifting her left hand to be in front of her torso, and then flicks it outward (iconically representative of her body pushing off from

the rock), with the maximum extension of the flick occurring precisely at the onset of the vowel in the word *rock*. During this, A, who has been looking away while telling the story, glances at B for around 700 milliseconds, then glances away again.

We'll turn now to B. Within 300 milliseconds of A turning to look at her, and while A is still completing her nod, B begins a cycle of nods. These continue for the entire time A is looking at her, and for a short time after A looks away again. While B is still nodding, but after A has looked away, B adds a spoken back-channel, *mhm*. This spoken back-channel begins just as A's speech segment is ending, during an optional adverbial (*a little bit*) but immediately after the core (and important) arguments of the clause (*I kicked off the rock*). B, as a listener, has been looking at A all this time, but shortly after A has finished glancing at her (and is not likely to glance towards her again soon), B takes a moment to glance away herself.

During B's glance away, says the word *but* in isolation, holding the floor and possibly planning her following speech. Then, as she prepares to shift to another event and another character in the story, she retracts her head back. As she finishes her third speech segment, she gives a small thrust of her head forward. This speech segment ends with more to be said (*and like..*), and B makes no response.

These four and a half seconds are not remarkably different the rest of the corpus they were drawn from. And yet they are remarkable in how much coordination between modalities they exhibit, and also in how many possible kinds of coordination they suggest. The temporal coordination between stressed syllables and peaks of manual and head gestures is well-reported in the gesture studies literature (e.g. Morrel-Samuels & Krauss 1992, de Ruiter 1998), and will be discussed in some detail in Section 5. The

tendency of head gestures to precede the speech they co-occur with (in B's back-channel) has also been reported on (Morrel-Samuels & Krauss 1992).

We can hypothesize about other kinds of coordination from this dialogue. For example, spoken back-channels often follow the offset of the speaker's speech, but here the back-channel precedes the offset. Is B's spoken back-channel timed to respond to the completion of the 'important' part of A' speech (the obligatory, core argumentation), or is it motivated by the combination of A's manual and head gestures? Do listeners give less feedback when speakers have signaled that there is more content to come? B's nod begins after A's nod begins. Is there within-modality responsiveness outside of speech – do nods beget other nods? B gazes away immediately following A's gaze-towards. If gaze is a signal of attention, does B wait until A's gaze-towards is complete (and tracks the average time between A's glances) to gaze away, so they will not be seen looking away? Or is B's gaze-away related to the fact that she just completed a back-channel? Is the retracting head gesture associated with narrative or perspective shifts? Does it tend to precede a jut-forward head gesture? Do nods temporally co-occur more with verbs or nouns?

In multimodal communication, there are presumably countless correlations and timing relationships that we still know nothing about and which existing models of language processing are not yet equipped to address. Given how young the fields of multimodal analysis and gesture studies are, we still know too little to know how to hypothesize about these interactions. These fields have so far been dominated by qualitative analysis of segments like Example 1. When approaching a system that is at once so familiar and so unknown, this is a useful approach, as it allows us to be guided by intuitions, which

are very often the beginning of good theories (although never the end). And close, qualitative analysis is a critical part of the analysis of systems as complex as verbal and non-verbal human communication.

Still, qualitative analysis suffers from some important weaknesses. First, it can lead to cherry-picking. The examples selected for a particular analysis may be rich, and be richly described, but the reader will be left not knowing the extent to which the analysis is generalizable to the broader communicative system. Second, without access to broad statistical tendencies of the phenomena being studied, a reliance on intuition can become an enemy, in the inevitable cases where our intuitions are wrong. Third, in young fields there are innumerable possible patterns still to find, and qualitative analysis is necessarily time-consuming.

Quantitative analysis is a useful complement to qualitative analysis, if the technology exists to allow it, and we are now reaching the point in multimodal analysis where it does. This dissertation is one example of how quantitative methods can be applied to multimodal data, with the goal of supporting qualitative analyses. In quantitative analysis, intuition can still be a treacherous guide, or a helpful one, so caution should always be exercised. But with a large, corpus-based analysis of a multimodal dataset, the findings from a qualitative analysis of one or more examples can be compared to the overall trends in the data, to show how representative the examples are. Even better, a well-coded dataset can be a resource for finding examples to analyze qualitatively. Analyses can be done much more quickly and comprehensively than with qualitative approaches, and the researcher can look at results from all sorts of relationships, even ones they never considered, to identify areas that warrant closer inspection.

It is this last point that I wish to focus on next. The behaviors involved in non-verbal communication may be interacting with each other and with speech in so many ways that we have not yet identified that it becomes hard to know best to approach the description. In the following section, I will make a case for using large-scale corpus analysis as a useful way of approaching analysis of non-verbal behavior, both as a way of identifying types of descriptive categories we can use, and identifying hypotheses we can test on other data, and with other methods.

2. The Difficulties Inherent in Describing Non-verbal Behavior

Verbal communication (language) can convey extremely detailed combinations of categorical information between two or more people, potentially across vast distances (via radio waves) and vast spans of time (via orthography). The units of language, whether morphemes, words, phrases, or constructions, display some regularities of form-to-meaning mappings across different speakers of a language, so that (barring effects of context or prosody) a given sequence of words can be produced by multiple people and reliably interpreted to mean the same thing. For example, if twenty people are asked to give the meaning of the word *bank*, we can expect fairly strong agreement in the semantics, the main differences being in which homonym they selected. This is not to say that verbal communication is perfectly regular and predictable, but it is quite regular and predicable when compared with non-verbal communication.

That said, verbal communication also displays a substantial amount of variation in the mapping of form to meaning. Much of the study of speech perception deals with this lack of invariance. This variation can occur at the lexical level (as with homonyms like *bank*), at the syntactic level (as in the multiple interpretations of *She saw the man with the*

binoculars), at the morphological level (the *-s* suffix in the word [bæŋks] can be one of three inflectional morphemes: possessive (*my bank's policy*), plural (*the banks are corrupt*), or third-person singular present (*he banks the ball off the backboard*), and is especially prevalent in the acoustics of speech. A single speaker's vowel spaces will often overlap with each other (Hillenbrand et al. 1995). In the Hillenbrand study (looking at the vowel space of English speakers from the American Midwest), some vowels, especially high back vowels, overlapped to the point where there was very little in the way of distinctive categories. Nevertheless, communication is only infrequently impeded by this variance, presumably because of contextual cues.

Non-verbal communication, which is made up of nearly every other communicative human behavior, can also encode a wide variety of semantic and pragmatic meanings, from spatial deixis or shape in manual gestures to emotional response in facial gestures to emphasis in nods and beat gestures. However, while some categories of non-verbal communication are highly conventionalized, like the 'thumbs up' and other emblems, most do not have clearly recognized form-to-meaning mapping seen in some parts of language, so common gestures like tilting one's head or turning a hand from palm-down to palm-up cannot be easily decoded without reference to the speech contexts they are produced in. Given this ambiguity, and the great variety of non-verbal forms, it is often unclear what the appropriate categories of non-verbal communication even are.

This leads to an important question: how should one approach describing a system of non-verbal communication when it is unclear what the units are doing, or even what the units are?

Take, for example, a head nod. If asked the meaning of a head nod, a North American might say it means “yes” or “agreement.” For some nods, this would seem correct. But nods are used in a variety of contexts where the notion of agreement seems out of place, such as a speaker nodding emphatically while recounting an event, a listener nodding as though to say “go on,” or a listener nodding to themself as the speaker’s meaning begins to dawn on them (see Poggi et al. 2010 for a more complete account). This might not be intractable – language also has polysemy, even homophony. But when we look closer at the kinematics of the nod head gesture, we see that there is enormous variability in the production of the gesture. A nod can vary on multiple continuous and categorical factors, such as magnitude, velocity, cyclicity, or direction, and these variations can potentially correspond to differences in meaning. This is a less tractable problem. The degree of variance in non-verbal communication is understudied, compared to the variance in verbal communication, and it remains unclear how similar or different these two systems are, in this respect. The findings in this dissertation may help shed some light on this question.

In addition to being composed of a set of uncertain forms mapped to a set of uncertain meanings, non-verbal communication is also complicated by the variety of its articulators. These include the eyes (both in terms of gaze-direction and blinks), the arms and hands, the head (both head gesture and facial gesture), and the entire body (both posture and proximity). Some of the parameters that these can vary on are similar across articulators, such as magnitude and duration, but most have their own degrees of freedom and affordances of articulation. Hands and arms obey the constraints of their joints and tendons, which still afford them an incalculable array of possible configurations. Spines

and shoulders are more constrained, but still allow for many different postures. Heads are even more constrained, limited to rotation and (for some) linear motion on three axes. Facial muscles allow for an extremely complex set of expressions. Eye blinks are much simpler, varying only on duration and rate, and eye-gaze varies only on the three-dimensional point of focus.

So non-verbal form-to-meaning mappings are uncertain, and the forms themselves are wildly different across multiple modalities. It is even unclear that "meaning" is a relevant property of some of these forms. And each of these, to a greater or lesser degree, somehow act in coordination with each other, and with speech, each time we communicate. This systematic coordination is quite extraordinary, and the cognitive systems that underlie it are still poorly understood. Most multimodal research done on the coordination between different modalities involved in communication has looked narrowly at two specific modalities within a speaker (often specific forms within each modality, such as head gestures and speech, e.g. Ishi et al. 2014), or at a single modality across two speakers (such as head gestures across speakers, e.g. Louwerse et al. 2012), rather than at the system as a whole (presuming, of course, that this is a system at all).

But this brings up another important feature of human communication. While communication now often occurs via computer or telephone, face-to-face interaction has been the most common context of communication throughout human evolution, and typically involves a wealth of both verbal and non-verbal information. When we are producing language, we are doing so with at least one other person as our target. This person, even if not an active turn-taker in the interaction, is consciously or unconsciously responding to signals from their interlocutor and giving signals in return, using all the

same articulators. The effect of these behaviors is not limited to the one producing them – it is, after all, communication. When planning what to say or how to respond, we are able to take into account all the available cues our interlocutors give us. A listener's frown can lead the speaker towards more precise lexical choices and a smile can inform the direction of a narration. Speakers and listeners co-construct the communicative act.

An attempt to describe the system of human communication, looking at it holistically, requires examining both the multimodal coordination within an individual communicator, and also the coordination across participants in the communicative interaction. We are still a long ways away from such a holistic model of communication, but we now have, or are developing the tools needed to do so. Besides being inevitable, there is a strong motivation for the field of linguistics for developing such a model. For many linguists, the goal of linguistics is to describe the relationships between the forms, meanings, and contexts of language. The multimodal coordinations of verbal and non-verbal behavior – which head gestures or facial expressions or gaze-shifts co-occur with which speech forms – are in fact a set of systematic contexts of language use[2]. For example, when Person A asks Person B "Do you love me?" and Person B responds "Of course," it is meaningfully different if Person B is looking at Person A while they respond, versus looking at Person C.

3. Corpus-based Hypothesis Formulation

We have established that the creation of more holistic models of human communication is both an important next step, and a difficult one. But how to approach the creation of a

[2] Or, from the perspective of multimodal analysis, these multimodal constructions are the forms themselves. When you have described enough of the environment, context becomes form.

model of such complexity? Current approaches, hypothesis-driven analyses of specific multimodal interactions, are a good place to start when most of the system is unknown – as more patterns are made clear, more pieces of the underlying structure of a model can be pieced together to form a systematic whole. But these approaches suffer from two problems: it is time-consuming to postulate them, and they are constrained by the kinds of hypotheses we are clever enough to make.

This dissertation proposes an alternative to hypothesis-driven testing of multimodal verbal and non-verbal behavior, or, rather, a complementary methodology, drawn from big-data approaches in other fields. Rather than formulating a hypothesis and testing it on a dataset using one or more methodologies, one selects methodologies, use them on every possible combination of behaviors in the dataset, and use the output to formulate hypotheses that can be tested at a future date, on future data. Looking through the results, one can look for correlations and other patterns that stand out, and perform more detailed qualitative analysis, compare these to findings in previous literature, or use them to formulate testable hypotheses.

This approach has a number of advantages, besides just the ability to quickly identify unthought-of patterns in a dataset. It can be very useful when applying methodologies to new kinds of datasets (as many multimodal datasets are) because you can compare effect sizes or likelihood measures across the entire dataset (although this is no substitute for independent replications). It can also be useful as a way of quickly comparing and contrasting two similar datasets – for example, corpora of two dialects, languages, or communicative contexts (e.g. narration and collaborative conversation). Or, within an

individual dataset, it can be used to identify patterns of differences across subgroups, such as genders or speech roles.

This approach also has a number of weaknesses. First of all, most of the analysis is entirely post-hoc (frequency distributions and durations of behaviors were all planned comparisons). In the cases where significance tests are run on the thousands of possible interactions between behaviors (Fisher's Exact test: Chapter 2, Section 6), there are no corrections for multiple comparisons, so it must be assumed that some proportion of the significant correlations are spurious, and this should lead to caution with regard to extrapolating from the results. Scientific claims about how human behavior works should be based on hypothesis-testing, and it would not be appropriate to use the original dataset as a place for this testing, meaning that a new corpus would have to be created to test any hypotheses on[3]. Another weakness is that, while many new and unexpected findings may arise, there will inevitably be a great deal of uninteresting results as well – behaviors one had no reason to think would interact, and which do not, in fact, interact. This may be an issue in this dissertation, as some effects can only be seen by comparing them to the surrounding absences of effects. I have tried to include only relevant figures in the body of the dissertation, but Appendix A will show these figures alongside a wider range of comparable figures where there appears to be no notable effects or correlations.

[3] One solution to this would be to run all the analyses on half of a dataset and test them on the other half. As this dissertation is not aimed at hypothesis-testing, and splitting the dataset would reduce the already small numbers of some sub-categories of behaviors, this will not be done here.

4. Existing Cognitive Models of Speech and Gesture

There is a large body of work dedicated to models of speech production and perception. The approaches to speech production are broken down into serial-processing models, which posit abstract modules of language processing that send planning information uni-directionally from conceptualization to articulation (Fromkin 1973, Garrett 1980, Levelt 1999), and parallel-processing models, which posit abstract nodes that connect semantic, phonological, morpho-syntactic, and lexical units together, where speech planning occurs bi-directionally via spreading activation (Dell 1986). As more holistic models are developed, these models will need to be integrated into the larger system of communication. Much of the early evidence for these theories came from research in speech errors, and there is more contemporary research showing that constraining manual gesture (by making participants sit on their hands) increases the rate of speech disfluencies (Rauscher, Krauss, & Chen 1996). Whether serial or parallel models are better equipped to integrate this kind of speech-external information is an open question.

While parallel-processing models better explain a great deal of linguistic phenomena, such as the higher than predicted rate of phonological speech errors that are real but semantically unrelated words, Levelt's model is the one that has been adopted and adapted by most gesture researchers. Levelt's model of speech production tracks speech from pre-linguistic concept to post-vocal tract sounds. It is composed of three broad stages or modules, each of which sends information uni-directionally to the next stage during speech production. The first stage, the Conceptualizer, contains conceptual and semantic information of words, and it is here that the pre-verbal message is generated. This message is passed to the Formulator, which encodes it into grammatically and

phonologically appropriate forms. This information is passed to the Articulator, where it is processed into motoric commands for the vocal tract, or other appropriate speech articulators. Finally, all this is monitored for errors, both auditorily by one's own speech comprehension system, and internally at various stages in production[4].

5. Gaps in Models of Speech-processing

Models like Dell's and Levelt's are admirable at explaining a large variety of linguistic phenomena, but they have not been designed to accommodate the variety of planning coordination we see between speech and other communicative modalities. There are three important aspects of this coordination that must be addressed by any model that integrates speech and multimodal processing: first, the temporal coordination of verbal and non-verbal behavior; second, the communicativeness of non-verbal behavior; and third, the way that non-verbal behavior can facilitate lexical retrieval and reduce disfluencies. Additionally, a model of communication that is truly holistic will take into account the interaction between speakers and listeners, building on the research into the co-construction of communication from linguistics, psychology, conversational analysis, and linguistic anthropology (e.g. Schegloff & Sacks 1973, Brennan & Clark 1996, Haviland 1977).

With regard to the first point, although there is much that we don't know about the timing relations between verbal and non-verbal behavior, some of these relations have already been described. As referenced earlier, iconic co-speech gestures and their lexical affiliates are well-synchronized, with the stroke of the gesture occurring just before or

[4] The research on models of speech perception is equally substantial, although this has not been a focus in gesture studies or multimodal analysis. However, complete models of interactive communication will need to account for both production and comprehension.

simultaneously with the lexical affiliate, but not after (Morrel-Samuels & Krauss 1992, de Ruiter 1998). We see similar temporal synchronies between other kinds of manual gesture and speech. Points of gestural prominence are produced simultaneously with points of prosodic prominence, often upwards of 90% of the time. What is meant by gestural prominence varies slightly from study to study, but refers broadly to the part of the gesture that is at the maximum distance covered, or otherwise most salient point. Loehr (2012) use "the peak of a [gestural] stroke," Leonard & Cummins (2009) use the point of maximum extension of a pointing gesture, while Roustan & Douhen (2010) and McClave (1994) use the downbeat of a beat gesture. The definitions of prosodic prominence vary slightly as well, from pitch accent (Loehr 2012, 2004), to stressed syllables (Shattuck-Hufnagel et al. 2007, to the dropping of the jaw during a stressed syllable (Rochet-Capellan et al. 2008). (See Wagner, Malisz, & Kopp 2014 for an overview of these timing relations.)

With regard to the second and third points, there is an ongoing debate in gesture studies as to whether gesture is produced primarily to communicate meaning, or to facilitate speech production, with supporting evidence being mixed. Supporters of the first theory claim that gestures provide information that is redundant or supplementary to information in the speech stream. Some support for this comes from learning research: for example, students show improved performance in foreign language-learning when words are accompanied by gestures (Kelly, McDevitt, & Esch 2009,), and children acquired mathematical concepts better when the teacher's speech was paired with non-redundant gestures (Singer & Goldin-Meadow 2005). Many other studies have tested whether gestures significantly contribute to communication, testing listeners' comprehension of

speech with and without gestures. A meta-analysis by Hostetter (2011) examined 63 such samples, finding that gestures did significantly contribute, and that this effect was increased when gestures were spatial rather than abstract, supplementary rather than redundant, and when listeners were younger. Additionally, in experimental settings where there is increased ambiguity in speech, listeners have been shown to pay increased attention to co-speech gesture (Thompson & Massaro 1986)[5].

Supporters of the other dominant theory, that gesture serves the purpose of facilitating speech production, suggest that the production of gestures that are semantically related to words (these words are called "lexical affiliates") aids working memory during lexical access of these words, particularly for words whose meaning can be also be represented gesturally (i.e. spatial meaning). This theory has been spearheaded by Robert Krauss, who has demonstrated that speakers who were prevented from gesturing spoke more slowly and produced filled pauses outside of expected syntactic junctures at a greater rate, which are suggestive of greater difficulty in lexical retrieval (Rauscher, Krauss, & Chen 1996). In a tip-of-the-tongue study, Pine, Bird, & Kirk (2007) showed that children were more accurate when they were not prevented from gesturing. It is well-known that speakers use co-speech gesture even in contexts when these gestures cannot be seen, such as while speaking on the telephone. Bavelas et al. (2008) showed that speakers gestured nearly as often while speaking on a telephone as did speakers in a face-to-face dialogue (although both groups gestured much more than the group giving a monologue to a tape recorder). Finally, research by Iverson & Goldin-Meadow (1997) shows that even

[5] It should be noted that some of studies show that gesture is *informative* (i.e. it may contribute information to the message being communicated) but not necessarily that it is intentionally *communicative*.

congenitally blind children, who have never seen co-speech gesture, will consistently produce gestures that resemble those of sighted children in form and content, suggesting a deep coupling between these two modalities.

These theories are not mutually incompatible, and it is likely that gestures can both be communicative and facilitate speech production. But each theory describes an integral relationship between speech and gesture, and future models will need to account for these relationships. The following section examines the current state of models integrating speech and gesture.

6. Integrated Models of Speech and Gesture

Some work has been done in the fields of gesture studies and multimodal analysis to connect what we know about speech processing with what we see in non-verbal communication. There are four primary models, several of which share the same core assumptions, and which I will discuss briefly here.

Growth Point Theory (McNeill 1992, McNeill & Duncan 2000): at the conceptualization stage in speech production, there is a single conceptual "seed" that sends planning information through both the speech processing system and the gesture processing system, leading to a multimodal utterance. This accounts for the fact that iconic gestures semantically relate to what is being spoken, and that the iconic gesture is temporally synchronous with its lexical affiliate (the semantically-related word that is co-produced with the gesture).

Sketch Model (de Ruiter 2000): an adaptation of Levelt's model. Similar to Growth Point theory, speech and gesture originate from the same point, and share the same

communicative function. In the conceptualization module, the communicative load is distributed across speech and gesture, and the Conceptualizer produces both a pre-verbal message and a Sketch, which is a pre-verbal message for the gesture production system. While the pre-verbal message is sent to the Formulator, to perform grammatical and phonological encoding, the Sketch is sent to the Gesture Planner, where it is encoded into gesture from the Gestuary, a sort of lexicon of gestural templates, and taking into account environmental factors, such as whether or not the person is holding a coffee cup. Finally, speech planning passes through the Articulator and gesture planning passes through Motor Control, resulting in overt speech and gesture. Speech and gesture planning are performed separately, but with occasional checks to ensure temporal synchronization.

Interface Model (Kita & Özyürek 2003): another adaptation of Levelt's model. This model was designed to account for the fact that the availability of linguistic information in a language influences gesture production. For example, in semantics, English speakers talking about swinging on swing are more likely to produce co-speech gestures that involve the arc of the swing's trajectory than Japanese speakers, who do not have verb "swing." In syntax, English speakers talking about a ball rolling down a hill (where the Path *down* and the Manner *rolling* are encoded in the same clause) are more likely to produce a single co-speech gesture combining these two elements, while Japanese and Turkish speakers, who express Path and Manner in two separate clauses, are more likely to produce separate co-speech gestures for each element. The Interface Model posits a single conceptual representation within which spatio-motoric information (to be sent to the gesture planner) is produced simultaneously in coordination with speech information.

Lexical Facilitation Model (Krauss, Chen, & Gottesman 2000): attempts to account for iconic co-speech gestures (called lexical gestures in the model), and also adapts Levelt's model of speech production. This is a speaker-oriented model, which makes the claim that lexical gestures are not produced to communicate information to an interlocutor, but to facilitate lexical retrieval. There is evidence for this, in that constraining co-speech gestures increases the rate of speech disfluencies (Morsella & Krauss 2004, and see Lucero, Zaharchuk, & Casasanto 2014 with regard to beat gestures), as well as evidence that gestures are produced even when there is no one to see them, both with blind children (Iverson & Goldin-Meadow 1997) and on the telephone (Bavelas 2007).

These models are good examples of theoretical, empirically-based approaches to modeling multimodal communication. Each is still limited, however, in accounting only for the integration of manual gesture (or even specific kinds of manual gesture) and speech. There is nothing to say these theories couldn't be extended to all other communicative modalities, e.g., that the Growth Point sends signals to coordinate gaze-shift, head tilts, and blinks, but very little of the work has been done to identify temporal and functional correlations between speech and these other modalities.

7. Modalities of Interest

This study looks at onsets and offsets of four different modalities, each of which has individually been demonstrated to contribute to communicative interaction. These four modalities are head gestures (involving rotation or linear motion of the entire head, not blinks or facial expressions), eye-gaze, speech, and manual gesture. More focus will be placed on the first three modalities than manual gesture because manual gesture has already received a great deal of attention with respect to its relationship to speech Manual

gestures are not parsed in this data set, although they were recorded with a Microsoft Kinect motion-capture system, which will not be included in this analysis. A brief synopsis of some of the research on some of these modalities will be given below.

7.1 Heads

The field of gesture research has taken manual gesture as its primary focus, but has not entirely ignored the co-speech activity of the head. Some early work has focused on categorizing and qualifying the forms of head gesture (Ekman & Friesen 1969, Kendon 1980, Duncan 1972, Schegloff 1987). More recently, research has examined the form, function, and timing of head gesture.

A number of researchers have worked on formally classifying head gestures (McClave 2000, Altorfer et al. 2000, Kousidis et al. 2013), which are typically described as having motion along axial rotation (roll, pitch, yaw) or on vertical lines of displacement (forward and backward motion, or side to side motion). These motions can be described in terms of their direction, speed, amplitude (including discernibility), or cyclicity (Hadar et al. 1984/5, Poggi et al. 2010, Wagner et al. 2014).

Researchers, particularly in the field of back-channels and conversational analysis, have also made attempts to describe the functions of head gestures, as back-channels (Poggi et al. 2010), as a means of claiming a turn (Maynard 1987), as a signal of a topic change, as having narrative and/or semantic functional properties, or with regard to their function in cognitive processing (Goodwin & Goodwin 1987).

With regard to timing, some research has studied the duration of speaker head gestures (House et al. 2001, Ishi et al 2014), and the temporal relationship between head

movement and speech (Dittman & Llewellyn 1968, Ishi et al. 2014, Hadar et al 1984), but much work remains to be done.

7.2 Gaze

Eye-gaze is an integral part of linguistic communication, given its importance in signaling attention. Indeed, there is more to eyes than just gaze, and researchers have examined aspects of eye behavior, like the communication of mental states (Baron-Cohen et al. 1997), the disambiguation of referring expressions (Hanna & Brennan 2007), or the communicative nature of eye blinks (Hömke et al. 2017). We focus on gaze because of its interactive capacity, which we will discuss below, and its potential relevance to gestural cues in the head and hands.

The formal properties of gaze are much less complex than head movements. The properties we look at in this study are the direction of the gaze, its duration, and the timing of its shift. In the context of conversational interaction, the direction is generally treated as a binary, either towards the interlocutor or away (Stivers et al. 2009).

In terms of its functional characteristics, gaze has been examined for many years by sign language researchers, as an interactive signal, as a syntactic cue for verb agreement (Neidle 2000, Thompson et al. 2006), and as a requisite element of comprehension. As a component of spoken languages, it has been assigned functions in cueing turn-taking (Goodwin 1980) as well as soliciting feedback from an interlocutor (Bavelas et al. 2002, Kendon 1967).

Recently, the timing of gaze shifts has received greater attention in linguistics research. It has been shown to be important in cueing turn-taking (Stivers et al 2009, Novick et al.

1996), as well as in establishing joint attention and common ground (Richardson et al. 2007).

7.3 Speech and Manual Gesture

The bulk of the research in gesture has focused on the hands, and the bulk of linguistic research has been on speech, and it would be impossible to attempt to summarize those findings here, the range of forms and functions of both modalities being too extensive to catalog (but see Section 5 for studies on the temporal coordination of manual gesture and speech). We do not examine these two modalities with the same level of detailed analysis as most other researchers do – our modalities of focus (head gesture and eye gaze) are under-represented in the literature relative to manual gesture and speech – and so here we limit our analysis to the timing of the onsets and offsets of manual gesture and speech – and we find there is still a great deal to be learned from such a simple, coarse-grained analysis.

That said, speech in this dataset has been coded according to certain functions that are thought to be relevant to interaction[6]. First, we distinguish between turns and back-channels. Back-channels are subdivided into different types, drawn from previous literature (see Section 8 for a discussion, and Chapter 2, Section 5 for the coding scheme). Turns are subdivided into categories based on both whether they are declarative or interrogative, and on clausal structure (complete vs. incomplete vs. filler). The motivation for the declarative / interrogative distinction was that interrogatives (coded here as grammatical questions and/or intonational appeals) have a function of soliciting

[6] Speech has also been transcribed and phonemically time-aligned, using the Montreal Forced-Aligner, but this data is not included in the current analysis.

some sort of response from the listener, which declaratives do not (or not as clearly), and we expected them to exhibit different co-occurrence patterns (which they did). The motivation for the distinction based on clausal structure was that responses to speech might be tied more to comprehension of content than to just speech itself, and that comprehension-based responses might be delayed until the core arguments of a clause had been completed.

There are many other kinds of categorical distinctions one could make about speech, using syntactic, semantic, or pragmatic categories. One could also incorporate prosodic information such as pitch contour or amplitude. A useful approach might be to included multiple categorizations, and look at the extent to which (for example) semantic and prosodic categorization schemes exhibit the same timing relations with other modalities.

8. Back-channeling

In doing an interactional, cross-modal analysis of communication, one feature that cannot be ignored is back-channeling, a behavior that has received substantial treatment in the fields of discourse and conversation analysis, but is frequently glossed over in treatments of speech production. Due to its significance in interactive communication, we briefly discuss it here.

Back-channel as a term has shifted from its original usage (Yngve 1970), where it referred to the channel of communication over which information was sent back from the listener to the speaker, without requiring the speaker to relinquish their turn. It has come now to mean any such signal that occurs on that channel, which are typically categorized as verbal or non-verbal. Back-channels are typically characterized as being communicative (Yngve 1970, Schegloff 1982, Krauss et al. 1977), although it remains

unclear precisely what is being communicated, whether anything is necessarily being communicated, or even what the defining characteristics of a back-channel are. It is an area that warrants more research.

The original (and still primary) focus of back-channel research has been in the speech modality. An assortment of communicative functions have been assigned to spoken back-channels – acknowledgments, continuers, affirmations, negations, assessments, newsmarkers, collaborative finishes, requests for clarification, and others – with varying degrees of overlap. They can be distinguish as phrasal (consisting of words, such as *woah* or *yeah*), non-lexical (consisting of meaningful non-word vocalizations, such as *hmm*), or substantive (consisting of more contentful, turn-like utterances, such as requests for clarification). Others have analyzed back-channels in other modalities, such as the head, face or posture changes (McClave 2000), following the lead from spoken back-channel research in terms of functions.

It is unclear, however, that a behavior taking the form of a back-channel is intended to signal anything back to the speaker. A nod, or a *hmm*, or even a *dude!*, may just as well arise from the actual processing of incoming information as from the desire to signal a response. One can easily imagine scenarios where either of these could be true: in one, a listener is paying no attention to the message of a speaker, but dutifully produces the necessary, well-formed back-channels at the right moments to make it seem like they are; in another, a listener is processing a difficult message from the speaker, and then nods and says *ahhh* as the significance dawns, with little or no thought towards what feedback the speaker may desire.

As mentioned earlier, it is also unclear where exactly to place the line between a back-channel and a turn. The distinction is usually made based on what the intent of the back-channeler is, whether that conforms to one of the pre-defined back-channel functions, such as continuer or assessment, or is something more substantive. This is problematic because it relies on a subjective interpretation of the back-channeler's intentions. But it is also problematic to attempt to define them based on whether or not they caused the interlocutor to relinquish a turn (Yngve 1970) originally posited the back-channel as being a kind of utterance that did not lead to a turn exchange), as this entirely ignores whatever the intent of the back-channeler might be. In conversation with frequent turn changes, this problem can become extremely thorny (which is in part why we have elected to use a story-telling paradigm). To be able to adequately fit the back-channel system into a larger model of interactional communication, what is needed is a formal analysis of back-channels, to describe their temporal and interactional behavioral characteristics. This study hopes to provide some useful data of this sort.

9. Methods

The methods used in this dissertation (to be discussed fully in Chapter 2, Section 5) are all related to timing relations within or across the four modalities of head gesture, eye-gaze, manual gesture, and speech, and within or across the roles of speaker and listener. They include simple analyses of durations and rates of behaviors, as well as co-occurrence patterns, likelihood measures, correlations, and frequency distributions. There are many other analyses that one can do, but timing relations seem like an excellent place to start, being foundational to the structure of a system that unfolds over time, and requiring very simple coding (merely the frames where a particular behavior is occurring,

or is beginning or ending). To understand how a complex system functions, we need to know how different components temporally coordinate with each other.

In the field of linguistics, there is already a great deal known about the timing relations and co-occurrence patterns of different language phenomena. In phonetics, we have identified trends in the co-occurrences of formant frequencies in common vowels, and trends of voicing distinctions within and across languages. In morpho-syntax, Greenberg and others have identified covariation between phenomena like word order and adpositional types, or trial and dual person marking – finding one pattern means another pattern is highly probable, though the inverse may not be the case (Greenberg 1963). Co-occurrence and dependency patterns of word classes are well-defined in phrase-structure rules, and the likelihoods of various syntactic alternations have been analyzed in depth based on a wide selection of co-occurrence patterns (Bresnan 2007, Arnold et al. 2000, Kendall 2011). Knowing these kinds of relationships allows us to think much more clearly about the underlying process involved in language processing and the causal relations between behaviors.

That said, there is still quite a lot we don't know about the how various properties of language co-vary with regard to timing. In phonetics, we know that pitch, amplitude, and duration covary, but we don't have measures of these covariances across languages and contexts. Between the subfields of syntax and semantics, we know that certain grammatical relations tend to co-occur with certain semantic roles, but we don't have quantitative, cross-linguistic measures of these associations, nor do we have cross-linguistic frequency distributions of grammatical features like (verbal or lexical) tense, aspect, and mood, and their interactions with each other.

When two or more components of a system pattern together, this points us towards the possibility of underlying causal factors. And getting a broad picture of what things do and don't correlate is especially helpful when exploring an understudied system like multimodal communication.

10. Some Issues Related to Description of Non-Verbal Behaviors

One remaining issue has to do with interpretation of this data, and this has two aspects. On the one hand, there is the issue of intentionality, and the on the other is the issue of causality. When looking at behaviors in a communicative setting like storytelling, it can be tempting to interpret any behavior as an intended behavior, as a cue meant to communicate something to the interlocutor. But it is important to remember that some behaviors can be intentionally *communicative*, and others can be merely unintentionally *informative* (at least, potentially informative, if the interlocutor is paying attention). Adaptors, a category of gestures that (typically) indicate discomfort, including tapping one's fingers or playing with the hem of one's clothing (Andersen 1999), are informative of one's mental state, but are not typically intentionally produced (in fact people often try to stop themselves when they discover they are doing them). When a listener shakes their head during the telling of a sorrowful event, this may be meant to communicate sympathy, or it may simply be an unintentional symptom of feeling sympathy. In the course of this dissertation, when interpreting multimodal patterns, I will offer suggestions of possible communicative functions, but it should be made clear that claims of communicativeness should be examined in qualitative analysis and/or experimental manipulation.

Similarly, the interactive nature of this kind of data makes it natural to look at speaker and listener behaviors as cues and responses. This is often no doubt the case, but one should still exercise caution in positing one behavior as the cause of another. Again, these are claims that should be tested in more controlled environments, and the approach used in this dissertation is a profitable way of forming these causal hypotheses.

11. Conclusion of Introduction

A comprehensive analysis of all possible timing relationships between all types of behaviors in a multimodal dataset makes it difficult to create an organization scheme for presenting results, because every potential organizing topic interacts with every other. Following the methods chapter (Chapter 2) and the chapter on summary statistics (Chapter 3), this dissertation will report the results of these analyses, broken down into four chapters: patterns within-role / within-mode (Chapter 4), within role / across-mode (Chapter 5), across-role / within-mode (Chapter 6), and across-role / across-mode (Chapter 7). Some methods will be more applicable in some chapters than others (for example, co-occurrence analyses are absent in chapter 4, because this chapter only deals with within-role, within-modality behaviors).

Within each chapter, results from each method will be presented for each modality, or combination of modalities. At the end of each chapter, results will be summarized and synthesized, and a set of possible hypotheses for possible future research will be presented, which I hope will be useful for useful for anyone interested in developing models of multimodal communication. But I hope the methods will prove equally useful for anyone analyzing similar data.

CHAPTER II: METHODOLOGY

1. Introduction

For the analyses and results in each chapter, the same scheme was used to code behaviors in the four modalities, and most chapters also use the same analyses. All data in this dissertation are drawn from a corpus of storytelling dyads recorded in the Discourse Lab at the University of Oregon in 2016. This chapter will describe the data collection process, the coding protocol, and the analyses used in the proceeding chapters.

2. Participants

Data for this experiment were collected from forty participants, all native speakers of American English. Participants were all University of Oregon students, and chose this experiment through the Psychology/Linguistics Human Subjects Pool, which awarded them research credit in an introductory Psychology or Linguistics class they were taking. Twenty-four were female, sixteen were male, ranging from 18 to 33 years of age, with a median age of 19. Participation was done in dyads, with eight female-female dyads, eight female-male, and four male-male. All included participants had not previously met. Four pairs of subjects were recorded but not included in this dataset (in addition to the forty participants), two pairs for being already acquainted, and two as a result of recording mishaps.

3. Procedure

On arrival, each participant was introduced to their dyad partner, and the experimenter asked each brief, informal questions about their studies, with the goal of providing a comfortable atmosphere. Participants were told that they would be engaging in conversation and storytelling with their partner, and that the experimenters were

interested in how people told stories. They were seated in armless wooden chairs, which were at 90 degree angles to each other, and at 45 degree angles from the recording cameras, and two meters away from the camera.

Participants were given five conversational prompts, followed by a storytelling prompt, which each took turns telling. These conversational prompts were designed to elicit various kinds of conversation, but for the purpose of this study, were mainly purposed to give participants time to become familiar with each other, both to facilitate natural and comfortable responses to the target storytelling prompts, and to allow each participant to observe and learn their partner's patterns of verbal and non-verbal behavior, particularly with regards to turn-taking. Since participants who knew each other previously were excluded, this ensured a level of equal familiarity for all participants. Participants were told to answer the prompts completely, but not feel pressured to continue if they thought they had finished. These were the five conversational prompts:

1. You are planning an itinerary for a mutual friend who is visiting Eugene for a weekend. What would you include in the itinerary?
2. You have a box of matches, a thumbtack and a candle. How would you use them to affix the candle to a wall so it doesn't drop wax onto the table below?
3. You have survived the zombie apocalypse and are looking for other survivors to join your group. What skills would you look for in potential members to help you survive?
4. You are planning the wedding reception for mutual friends. All the bride's family are very liberal, and all the groom's family are very conservative. What would your reception look like?

5. Your roommate has stopped cleaning and doing dishes, and instead just sits around all day watching television. They recently went through a break up and may be depressed. How would you deal with the situation?

Responding to the conversational prompts took around 15 to 20 minutes. Based on analyses from pilot experiments, this amount of time was (subjectively) judged to be sufficient for participants to accustom themselves to each other's styles. Following the conversational portion, participants were asked to think of a personal story according to the following prompt:

6. Tell your partner a story about an experience you had where you were afraid you might die.

The corpus analyzed in this dissertation is entirely drawn from responses to this prompt. The prompt is drawn from similar questions used in sociolinguistic interviews (Labov 1972) and has been used successfully in experimental gesture studies (Bavelas 2002). It is especially useful in a laboratory context, where elicitation of speech can often be forced or overly self-aware, leading to lowered rates of gesture compared to naturalistic speech contexts. However, because the both storytellers and story-listeners tend to become quite emotionally involved in stories about near-death, this corpus is quite rich in behavior across all our coded modalities. As one measure of how naturalistic these stories were, only one participant looked at the camera during the elicitation, and his storytelling seemed otherwise quite comfortable and natural.

In order to accommodate to participants who either couldn't think of such a story, or didn't feel comfortable telling one, the experimenter explained that they could

alternatively tell a personal story about a thrilling or embarrassing event in their life. Three participants (in three separate dyads) opted for this alternative, all three opting to tell an embarrassing story. Extensive comparisons have been done on near-death and alternative stories, and no distinctive categorical differences have been found, so data from all stories have been collapsed in this study. Participants took turns telling these stories. Comparisons were also done across groups that told their story first, and that told their story second, on the possibility that the latter storyteller would differ due to greater familiarization with their partner. No differences were found across these groups, either.

The procedure went as follows: the experimenter gave a conversational or storytelling prompt, began the recording, left the room, and waited outside the room, next to a window (that did not allow them to see the participants). When participants had completed responding to the prompt, they signaled the experimenter by throwing a cardboard box past the window. They were told that if they wished, they could try to land the box on the seat of a chair just past the window (about three and a half meters away), and were told the success rate of previous participants, and indirectly encouraged to try to beat them. Then the experimenter re-entered the room, ended the recording, asked for a brief synopsis of their response, offering approving feedback, and started the next prompt.

During debriefing, participants were asked what they thought the actual goal of the experimenters was. Many guessed that we were interested effects of culture, age, or sex on storytelling styles. Only two asked whether it had to do with manual gestures, and these two did not seem to gesture differently than other participants, based on rates and durations of production in the four modalities. No participants guessed that head gesture

and eye gaze were of interest. Three participants asked if box-throwing accuracy was the real goal of the experiment. It was not.

4. Recording

Participants' audio and video were recorded at 30 frames per second with a Canon XA10 HD and two Rode NT1-A directional microphones, as well as with Microsoft Kinect (v2). Kinect data was judged sufficient for its temporal resolution, but spatial resolution was judged inadequate for the current analysis, compared to non-automated coding.

5. Coding

Video and audio were imported into ELAN for coding (Wittenburg et al. 2006). For each of the four modalities in question, and for both storyteller and story-listener, onsets and offsets were hand-coded. This coding differs slightly for each modality. For head gesture, an onset is whenever a participant begins to move their head, either from motionlessness or on a new axis of motion, and an offset is when they stop moving their head for at least 250ms, or when they begin to move their head on a new axis. For eye gaze, an onset is when a participant's eyes begin looking at the interlocutor's face, and an offset is when they begin looking away (blinks and gaze-shifts shorter than 500ms were not coded). For manual gesture, an onset is whenever a participant begins to move one or both hands, and an offset is when they return their hands to rest position, or cease to move their hands for at least 500ms (effectively defining a new rest position). For vocalizations, an onset is whenever a participant begins to make a sound with their mouth, and an offset is whenever they stop for at least 150ms (following Redford 2013).

The intention in coding was to account for behaviors that we believed might be involved in the multimodal system of interactional communication. There are several kinds of

activity that occur in these modalities which are not thought to be implicated in interactional communication, and which were not coded nor included in the analyses. In the modality of head gesture, we did not code head motion resulting from sneezes, coughs, or self-touching with the hands. In manual gesture, we did not code self-touching hand motion, such as adjusting clothes or hair, or repetitive hand tics, such as tapping one's leg or twiddling one's thumbs. It is possible that some of these behaviors may be communicative, and almost certainly that most of them could be informative, but the information they convey is outside the scope of this thesis.

5.1 Heads

In addition to the onsets and offsets of head behaviors, we also coded each head gesture according to several categorical features. A summary of these behaviors is shown in Table 1.

Table 1. Head gesture coding scheme

	Pitch (x-axis)	*Yaw (y-axis)*	*Roll (z-axis)*	*Linear displacement*	
	Nod	**Shake**	**Tilt**	**Jut**	**Retraction**
Half cycle	nod up nod down	turn away turn towards	tilt away tilt towards	jut	retract back
Single cycle	single nod	single shake	single wag tilt + return	single jut	single retraction
Repeated cycle	multiple nod	multiple shake	multiple wag	multiple jut	multiple retraction

Compared to manual gesture, head gesture is relatively constrained in the possible forms that can be performed[7]. The head is a ball, sitting atop flexible stalk. It can rotate on three axes. Pitch is the axis that pierces through the ears (this is comprised of *nods*), yaw is the axis that pierces through the top and bottom of the skull (this is comprised of *turns* and *shakes*), and roll is the axis that pierces right between the eyes (this is comprised of *tilts* and *wags*). (In the future, these three axes will be referred to using the Euclidean coordinate system: pitch is the x-axis, yaw is the y-axis, and roll is the z-axis.) It can also, to a limited extent, move linearly on top of the spine. This is never seen on the y-axis (moving straight up or down from the spine), but it is seen on the other two axes. Linear displacement on the z-axis, towards and away from the interlocutor, is quite frequent, and is coded here as *juts* and *retractions*, depending on the direction. Linear displacement on the x-axis, the side-to-side head slide, is attested, but was only seen in one participant in this dataset, and so has not been coded.

Given the location of the head on top of the spine, there is substantial axial head motion that can occur which is dependent on motion of the rest of the body. For example, unless otherwise prevented, the head will move when a participant leans forward, turns their body to the side, or leans to the side. This dependent motion seems different in kind from independent axial head motion. For this reason, head motion was always coded relative to the position of the shoulders.

[7] There is a question to be raised about how quickly the head must be moving to be considered a head gesture, which had implications for inter-coder reliability. If the head is lowering very gradually over time, it seems difficult to call this a nod, but the line between gradual motion and gesture is not clear. Moreover, there are some participants whose heads are almost constantly in slight, barely detectable motion. The coder had to rely on subjective analysis to determine whether a particular head motion qualified as head gesture.

In addition to axis of motion, head gestures differ in their cyclicity. Whether their motion is rotatory or linear, they can be broken down into three categories: a half cycle (motion in one direction, with no immediate return), a single cycle (motion in one direction followed by a return to the original position), or multiple cycles (multiples successive iterations of a single cycle). Half cycles can occur in either of two directions, which are coded as *up* or *down* for nods, *towards* or *away* for tilts and turns, and *juts* or *retractions* for linear motion on the z axis. Single cycles also begin in one of these directions, but we have not coded rotatory head gestures that begin in different directions as distinct categories (i.e. a *single nod* may begin with upward or downward motion), as we did not see them behaving in distinct patterns in the same way some half-cycles did. However, single cycles of *juts* and *retractions* were coded distinctly, as they seem to pattern similarly to their corresponding half-cycles. Multiple cycles of head gestures follow this pattern as well. They involve two or more repeated cycles of motion, with no cessation of movement for more than 250ms.

Sometimes a head gesture will involve motion on more than one axis or line of displacement. For example, a single nod might coincide with a jutting forward of the head, or a cycle of head shakes might coincide with a slow single nod. When this occurs, the gesture is coded for all relevant categories. It is an interesting question whether such composite head gestures have conventional meaning distinct from those of their component axes, but unfortunately this analysis was not possible with this data set. Composite head gestures make up fewer than 10% of the total number of head gestures, and their composite nature gives us too many types and too few tokens to do meaningful analysis. We have excluded them from this analysis.

In creating this coding scheme, our goal has been to be kinematically systematic, basing our categorical distinctions on simple parameters of kinetic motion. We are trying to avoid making claims about functional similarities of any of the categories, and so have tried to avoid collapsing them with each other. However, for some categories (*multiple wags*, *multiple juts*, *multiple retractions*), there were too few tokens (N < 5 in each case) to judge any individual characteristics, and so we have collapsed them with *single wags*, *single juts*, and *single retractions*.

With regard to rotation on the y-axis (*shakes* and *turns*), we have elected to not include head turns in our analyses. Turning the head away from the interlocutor largely coincides with shifting one's gaze away, and turning the head towards largely coincides with a gaze shift towards. Of these two behaviors, it seems likely that the gaze shift is the more salient to the interlocutor. Analyses including both gaze shift and head turns show that head turns make little independent contribution to the interaction.

With regard to rotation on the z-axis (*tilts* and *wags*), tilts were coded with respect to the interlocutor, rather than the tilter, so a tilt is coded as towards or away from an interlocutor, rather than to the left or right. In our recording context, participants sit a 90 degree angle from each other, and so the towards/away distinction results in falling more on the z-axis than the x-axis. Since a major focus of this study is interaction, this seemed to be the more relevant distinction. However, there are arguments for other coding schemes, and in a context where participants are directly facing each other, a left/right categorization might make more sense (although, given the near symmetry of the human body, it's not clear what a left/right tilt distinction might indicate if not toward or away from another entity).

Finally, also with regard to the z-axis, we made a categorical distinction between a single wag and a tilt + return. Both of these involve a single cycle of rotation on the z-axis, but most people view the two gestures as having categorically different functions. They can also be distinguished by their forms: in a tilt + return, the motion is purely on the z-axis, and the return typically doesn't extend past the original point of departure, while in a wag, there tends to be a small amount of rotation on the y-axis, and on the return the head extends slightly past the point of departure before returning to its original position, resulting in something like a lopsided figure eight path. Single nods and shakes also often follow this slight overshot on the return.

Head gestures are a strong focus in this dissertation, and their coding scheme was the most complex, involving multiple kinematic dimensions. Inter-rater reliability was measured with another coder, who was trained on the coding scheme detailed above. The main source of differences between the coders had to do with what velocity of head movement constituted a head gesture, and were mostly limited to whether or not a half-cycle was considered to be a gesture or not. For example, a very slow rise of the head might be coded as a nod-up by one coder and not be coded as a gesture by the other. The decision was made to be conservative in the direction of identifiability – the velocity of movement that was judged to be a gesture by both coders was the velocity used to code other gestures.

When these slower half-cycles were removed, inter-rater reliability was assessed using Cohen's Kappa. This was done in two ways. The first measure looked at whether or not each coder's head gestures occurred at the same time and were coded as belonging to the same category, which resulted in a Kappa of 0.85, demonstrating strong agreement in

category assessment[8]. The second measure also looked at how closely the onsets of each pair of coded head gestures matched, with distances above a certain threshold being considered a category mismatch. Inter-rater reliability was assessed at increasing threshold of distance. At 100ms, Kappa = 0.34, at 300ms, Kappa = 0.72, and at 500ms, Kappa = 0.85. On the one hand, we should insert some caution into our analysis of these boundaries, and on the other we should consider that individual variation in the detectability of head gestures probably also extends to speakers and listeners, and that some head gestures may have clearer forms and functions than others.

5.2 Gaze

Gaze, along with manual gesture, has the simplest coding of all the modalities in this dataset. It is binary: a participant is either looking directly at their interlocutor, or they are not. There are a number of finer distinctions one could make in coding gaze direction, which may well be relevant in the system of communicative interaction. Participants look up, down, and to the side. They look away from their interlocutor's body, or at it, or specifically at their gesturing hands. When they shift their gaze away from their interlocutor, it may be only a short distance, so a head or manual gesture is still visible in their peripheral vision. Any of these finer distinctions in gaze direction may be communicative or informative, and may be involved in communicative decisions made by their conversation partner – gazing up may be conventionally understood to mean a person is considering the content of the narrative, or knowing that one's head can be seen peripherally may influence the likelihood of a head gesture. These categories, however, are not reliably detectable from our video data. The binary distinction, of gazing towards

[8] For cases where there was disagreement, there was typically still agreement in axis of motion.

and away from the face, is highly detectable, and proves to be one of the most predictive distinctions in our data set. Plus, humans are particularly sensitive to whether or not another person's gaze is directed toward them or not.

Another coding decision involved how long a person must break eye-contact to be considered gazing away. All participants blink, usually for no more than 200ms. Some participants blink for extended periods of time, even upwards of 1000ms. Other participants might glance away from the face for a few hundred milliseconds. The question is whether these behaviors constitute gaze shift that is in some sense 'meaningful' for the interlocutor?

Following Bavelas (2002), we see mutual gaze towards the interlocutor as a 'gaze window,' during which visible cues such as head and manual gestures can be seen. It may also be worth thinking of gaze away from the interlocutor as a window of a different sort, a window in which there is less urgency to provide visible (or even possibly aural) signs of attention, and so there is time to relax, process, and attend to different parts of one's interlocutor. For this window to be meaningful, though, it must be long enough to ascertain that it exists. The amount of time required to engage muscle response to visual stimuli takes is at least 150ms, and substantially more time for conscious responses, and initiating a physical response takes about the same amount of time (Libet 1985). This means that most blinks, and some brief glances away, aren't even fully processed before they're over. For this reason, we excluded them from our coding scheme.

5.3 Speech

Onsets and offsets of all vocalizations were coded. Following Redford (2013), we treated any cessation of vocalization longer than 150ms (and longer for certain fricative-final

words) as a pause, and coded an offset. Vocalizations could be speech or non-speech, but not breathing. Non-speech vocalizations were predominantly laughs, but also included a small number of sighs and clearings of the throat (too few of these last to do any meaningful analysis). Laughs were coded as a single category, although it seems likely that there are many distinct forms of laughter, which may convey different information and elicit different responses. However, such an analysis would have to be a separate project.

Speech vocalizations were transcribed and time-aligned at the phoneme level, using the Montreal Forced Aligner (McAuliffe et al. 2017). Speech was also categorized according to two categories of function: turns/back-channels and mood.

In the turn/back-channel categorization, we classified each speech segment as being one of a set of back-channel functions, or as a speech turn. As discussed in Chapter 1, Section 8, there is no clear method for distinguishing between a speech turn and a spoken back-channel. Since back-channels are optional and thus largely unpredictable, they cannot be easily placed in an adjacency pair. The traditional method of classifying turns as speech utterances that succeed in taking or holding the floor only takes into account the outcome, not the intention of the utterance, which fails to capture its expressive functionality.

Instead of using taking the floor as a diagnostic, we have first looked at whether a speech segment is responding to the interlocutor's speech. If it is, we look to see whether it falls into one of eight back-channel categories: acknowledgements, affirmations, assessments, collaborative finishes, continuers, negations, and newsmarkers. All other speech segments are classified as speech turns. These back-channel categories are drawn from

categories developed by multiple researchers in back-channels (Schegloff 1982, Gardner 2001, Drummond & Hopper 1993, Sacks et al. 1974). They are defined as follows:

- Acknowledgements express an acknowledgement of the content of the interlocutor's speech. Frequent examples include *yeah* and *mm*.
- Affirmations indicate agreement or support of something the interlocutor has said. Frequent examples include *right* and *yeah*.
- Assessments express some emotional or affective response. Frequent examples include *oh my god*, *no way*, or *cool*.
- Collaborative finishes occur when a listener joins a speaker in finishing their own utterance.
- Continuers encourage the interlocutor to continue speaking. Frequent examples include *uh huh* and *mhm*.
- Negations express a polar negative response to a question or other appeal from the interlocutor. These are not frequent in this data set, but *no* and *uh uh* are examples.
- Newsmarkers indicate that content from the interlocutor is new information. Frequent examples include *oh* and *really*.

This coding scheme suffers from some flaws. We cannot read participants' minds, and so can never be certain what a given back-channel is meant to express. There is also likely to be overlap across these categories – newsmarkers may also involve some amount of affective assessment, and many back-channels may be intended or interpreted as indicating acknowledgement, or that the interlocutor should continue. One approach might be to assign back-channel functionality based on the lexical items themselves, but

there are cases where a single word, such as *yeah*, can be used to indicate either acknowledgement or affirmation, with the two functions being disambiguated by prosodic features of the word.

Speech turns are also subcategorized according to their function. The main categorical distinction coded in speech turns is that of interrogative, declarative, filler, and incomplete. An interrogative speech segment ends in some appeal to the interlocutor, marked in the prosody, the syntax, or both. Prosodically-marked interrogatives end with raised pitch, and may be declarative questions (e.g. *You know what I mean?*) or merely statements with question intonation (e.g. *So like because I was near the tree I grabbed it?*). Syntactically-marked interrogatives include yes-no questions and *wh*-questions, whether or not they also have question intonation. A declarative speech segment is not marked syntactically or prosodically as interrogative, and contains the predication of the clause. Often a clause will be broken into multiple speech segments by multiple pauses. In such cases, connectives (e.g. *so then*, *and uh*, or *but like*) were classified as fillers, along with filled pauses (e.g. *um*, *uh*, or *well*). Content-filled parts of the clause that had not yet expressed the predication (whether that predication turned out to be declarative or interrogative) were classified as incomplete. For example, in the clause below, 1a is classified as incomplete, and 1b is classified as declarative.

1a) *So I was..* (incomplete)

1b) *..three at the time, three or four.* (declarative)

5.4 Hands

Like gaze, manual gesture is coded as binary, being present or absent. It was not coded to the same level of detail as has been done in many gesture studies analyses. This was due in part to a desire to use automated coding of gestures (participants were recorded with the Microsoft Kinect (v2) in addition to regular video, but this data is not included in these analyses), but also to shift greater focus onto head gestures, which haven't received nearly the same degree of attention. Hand motion that had some functional purpose related to self-grooming, itching, adjusting clothes, or other self-touching was excluded. Adaptors, such as repetitive hand tics, tapping one's fingers on one's legs, twiddling thumbs, or repetitive leg scratching, were also excluded. The remaining hand motions were coded as gestures. Onsets of manual gesture were coded any time one or more hands transitioned from motionlessness to motion, and offsets were coded whenever both hands remained motionless for at least 500ms. The exception to this was case where hands returned to rest position (since they were seated, most participants' hands rested in their laps, and a few let them hang straight down from their shoulders). Returning to rest position was coded as an offset, even if the hands immediately left again, and leaving rest position was coded as an onset, even if the hands hadn't stopped moving for 500ms.

6. Analyses

To explore the timing relations in these multimodal data, a number of analyses were run on each behavior and, in many cases, on each relationship between the different behaviors. These include summary statistics of durations and rates of behavior segments, conditional probabilities, odds-ratios, and frequency distributions of individual and co-

occurring behaviors. Since these are largely the same across each proceeding chapter, a brief summary of the analyses will be given here.

To get a sense of the timing characteristics of individual behaviors, durations of each behavior (at the level of modality and by subtype) were calculated, from onsets to offsets of behaviors. In some cases, durations of 'non-behaviors' were also calculated, looking at the time-lag between the end of one behavior segment and the beginning of the next one, like pauses between speech or periods of gaze-away. Rates of behaviors were calculated from the entire corpus, and correlations of rates were calculated across speakers and listeners participating in the same stories.

One important characteristic of timing relations between behaviors is the degree of co-occurrence, either in terms of the amount of overlap between behaviors or the number of instances where two behavior boundaries occur near each other. This can give us a sense of the dependencies (i.e., temporal dependencies, not necessarily causal) between two or more behaviors. These have been calculated using *conditional probability*. For each frame in the corpus, each type and subtype of behaviors is coded as being present or absent. The conditional probability of each pair of behaviors is determined by counting the number of frames in which behaviors A and B are both present, and calculating this as a proportion of the total number of frames that behaviors A and B are present individually. For conditional probabilities of proximal behavior boundaries, first a span of frames is chosen, then the number of instances of behavior boundary A are counted that occur within this span of behavior boundary B. This is taken as a proportion of the total number of behavior boundary A's.

The degree of co-occurrence between two behaviors is useful for describing the temporal relations between elements of a system, but when different behaviors have different frequencies, it is also helpful to have a measure of how much this amount of co-occurrence deviates from random chance. To measure this, we calculate the expected amount of overlap between two behaviors (from the total number of frames of each behavior and the total number of frames in the corpus) and compare this to the observed amount of overlap using Fisher's Exact Test. The p-value from this tells us whether the observed and expected overlaps are significantly different from each other. Equally helpful is another test statistic, which is an *odds ratio* of the two overlaps. This is a bi-directional measure of association, telling us how much more or less likely the observed overlap is than we would have expected from random chance given the overall frequency of each behavior. This is a particularly useful measure when examining large amounts of data to see what stands out as strongly or weakly associated, and for identifying comparative trends across categories.

In some cases, such as looking at the frequency of co-occurrence of two behavior onsets within a certain moving window of time, Fisher's Exact Test is not applicable, because the expected distribution is calculated from the total number of counts, and these cannot meaningfully be compared to counts within a moving window. In these cases, we will instead rely on a measure of association called symmetric conditional probability. This measure is simple to calculate, as it is the product of each of the one-way conditional probabilities of the two behaviors in question. It does not tell us whether the interaction between the two behaviors is different from chance, much less in which direction, but it does give a sense of the mutual dependence between the two behaviors.

To examine sequential patterns, *n-grams* of behaviors were also calculated, within and across role, and within and across behaviors. N-grams, like conditional probabilities and odds-ratios, are common in corpus linguistics. However, the nature of a text corpus differs importantly from a multimodal corpus. In text corpora, there is usually a single stream of units, and n-grams are repeated sequences in this stream. In multimodal corpora, the units are messier. The behaviors themselves occur on different streams and are of widely differing durations, so they can overlap with each other as well as precede and follow. Using onsets and offsets as n-gram units makes this easier, but with multimodal data there is still the possibility that two such boundaries will occur at the same time. Additionally, because the distance between consecutive multimodal behaviors is not fixed (as is the case between two consecutive words in a traditional corpus), there is the issue of determining how near two behaviors must be to consider them as part of a sequence of behaviors. Two methods are used here: looking at sequences of behavior boundaries that occur within a set span (such as five seconds) and looking at sequences of behaviors that are within a set span of each other.

Finally, frequency distributions of behaviors can be useful for capturing patterns of behaviors over time. I use these in two ways. To look at the distribution of individual behaviors over the time-course of a story, I use '*story histograms*.' In these, each of the forty stories is compressed to the same number of frames (1000), and a histogram is shown of the frequency distribution of a particular behavior. These distributions often display patterns that relate to the narrative structure of the discourse, increasing or diminishing in frequency as the story progresses through rising action, climax, and falling action. To look at the interactions between behaviors, I use '*window histograms*.' In

these, all the instances of behavior boundary A that occur within a particular time-window of behavior boundary B are plotted in a histogram. (Sometimes there may be more than one instance of behavior boundary A, and in these cases only the A boundary nearest to the B boundary are plotted.) Any peak in these histograms indicates a tendency for behavior boundary A to cluster near behavior boundary B. The skew of the distribution tells us something about whether behavior A tends to precede, follow, or occur simultaneously with behavior B. The slope of the peak informs us about the strength of the timing relation. There are a wide variety of behaviors being examined in this dissertation, some more frequent than others. Because of this disparity in frequencies, the bin size is not kept constant across all histograms, but instead the *doane* algorithm (the default in Python's matplotlib library) is used to calculate the optimal number of bins.

CHAPTER III: SUMMARY STATISTICS

1. Introduction

While chapters 4 through 7 deal with the interactions and sequential patterns of the behaviors of interest, this chapter provides an overview of a number of summary statistics of the behaviors and stories. Section 2 describes the kinds of content that the stories were about. Section 3 covers the durations and rates of each behavior at the broad level of modality. Section 4 covers rates of subtypes of head behavior, broken down by axis and cyclicity of motion, while Section 5 covers the durations of head behaviors. Section 6 covers both rates and durations of speech types, looking at subtypes of speech turns and spoken back-channels.

2. Overview of the Stories

The personal near-death stories that these data come from belong to a particular kind of human interaction, one that is probably universal across cultures. Personal story-telling is surely universal, and while not everyone may have a story of being near death, people are drawn to tell stories about events that are exciting or thrilling for one reason or another. In emotionally-charged narratives like this, we become invested in the story while telling it, and we have reason to believe our listener will be invested as well (although this is not guaranteed, and we may still have to work to keep their interest). These stories may be different in kind from other storytelling contexts, and are certainly different from interactional contexts like arguments, problem-solving, or instruction-giving. To give a sense of the kinds of topics that were being discussed in these data, Table 2 breaks down the stories into a set of broad categories.

Table 2. Topics of near-death stories

Topic	Frequency
Vehicle (collision)	8
Vehicle (near-collision)	8
Risk of drowning	8
Injury (non-vehicular)	4
Medical condition	3
Animal	3
Assault	2
Lost	1
Embarrassing	3

The largest portion of the stories had to do with vehicle-related danger: 40% of participants told stories about near-death experiences involving vehicles, and half of these involved a collision. Near-collision stories often involved a car turning into dangerous traffic, hydroplaning or otherwise losing control, and narrowly avoiding an accident, although in one case the near miss was a trio of tornadoes approaching the highway. Collisions involved contact with another vehicle or a stationary object like a ditch or a house. These particular statistics are somewhat unsurprising, as motor vehicle-related accidents are the leading cause of deaths for 5 to 24-year olds in the US (NCHS/WISQARS, 2016). Risk of drowning was another common source of danger. These ranged from stories of falling into rivers or lakes as young children to craggy reefs while snorkeling and oxygen mishaps while scuba-diving. Injury-related stories involved major breakage of skin and/or bones, such as crashing into rocks on a rope swing or being hit in the stomach with a flying blade while cutting rebar. Medical conditions

included collapsed lungs and hypothermia. Animals involved horses and bears and, in one case, a dog that wrapped its leash around its owner, resulting in a concussion. There were only two instances of violent assault: a knife attack at a bar and a robbery at gunpoint in Brazil. There was one instance of a participant being lost, in New York as a child. Three participants chose to tell an embarrassing story, either because they couldn't think of a near-death experience to tell, or didn't wish to. These included stories of slipping on ice in front of a school bus of children and of dangling by one's snow-pants from a moving T-bar at a ski resort. Qualitatively, these three embarrassing stories still engendered the same kind of interest from the listeners. Quantitatively, in a set of comparisons with near-death stories for each of the methodologies used in this dissertation, they did not differ from the near-death stories. Most categories were equally well-represented across sexes, although women had a higher ratio of vehicle (collision) to vehicle (near-collision) stories, and all three participants who chose not to tell near-death stories were women.

3. Duration and Rates of Behaviors by Modality

We look next at how behaviors in the four modalities of interest are distributed across the roles of listener and speaker. Table 3 shows summary statistics of these distributions at a very broad level, the level of the modalities themselves, comparing listeners to speakers.

Table 3. Summary statistics of behaviors by modality and role (ms.) (Non-role behaviors are excluded)

	Freq.		Duration				
	Count	**Rate (per min.)**	**Mean**	**SD**	**Median**	**Min**	**Max**
Listener Head	1151	12.05	793.72	513.38	667	100	2867
Listener Gaze-towards	441	4.62	8054.81	10834.93	3967	67	66667
Listener Hands	69	0.72	1282.62	786.58	1000	367	3700
Listener Speech	719	7.53	556.29	326.52	433	67	1833
	Count	**Rate (per min.)**	**Mean**	**Std**	**Median**	**Min**	**Max**
Speaker Head	2305	24.14	653.36	399.89	533	100	2200
Speaker Gaze-towards	967	10.13	1647.18	1262.59	1300	100	6867
Speaker Hands	898	9.40	2308.95	1966.39	1600	233	10833
Speaker Speech	1866	19.54	2143.84	1586.93	1733	100	7367

One clear difference across roles is that speakers are simply *doing more*. They outpace listeners in each modality by a ratio of around two or three to one, except in manual gesture, for which they outpace listeners at a ratio of fourteen to one. For speech, this is not surprising, given that speakers are providing the bulk of the information (and, it is generally assumed, controlling the flow of the conversation). Speakers produce speech segments at a rate almost three times that of listeners (once every three seconds compared to once every eight seconds), and these segments are on average much longer (2144ms compared to 556ms), a significant difference ($t(2583) = 26.61$, $p < .001$). Speech segment durations for both roles are negatively skewed, with the majority of the segments being of shorter duration than the mean.

Speaker head gestures occur at twice the rate of listener head gestures, with speakers gesturing nearly once every two and a half seconds, and listeners once every five

seconds. These gestures are not, however, evenly spaced, as we will see in (Chapter 4, Section 2). Listener head gestures, unlike listener speech segments, tend to be significantly longer than speaker head gestures ($t(3454) = 8.82$, $p < .001$). This results in part from speakers shifting more rapidly from one head gesture to another, and from listeners preferring longer cyclical head gestures on the same axis, as we will see shortly. Overall, however, duration of head gesture is the most similar of all across-role modalities. This may be from physical constraints, or even from functional constraints. If we consider each modality as having a set of forms and functions it can articulate, speech having all the complexity of a language's grammar and lexicon, and manual gesture having all the complexity of freedom provided by shoulder, elbow, wrist, and finger motion, then heads seem rather impoverished in their complexity[9], which may help explain the relative similarities in duration across roles. They can rotate on three axes and shift linearly to a limited degree on two axes, forward and side-to-side. Whether these limitations of form correspond to limitations of meaning remains unclear. We often think of head gestures as having only a small number of conventionalized functions, such as affirmation for nods and negation for shakes. But head gestures can express a wealth of functions, particularly in context with other multimodal behavior. This will be taken up in Chapters 5 through 7.

The distribution of manual gesture across roles is one of the most surprising findings in this table. Given what we saw in speech, it might not be surprising that speaker manual gesture segments are significantly longer than those of listeners ($t(965) = 4.31$, $p < .001$),

[9] Gaze, while highly salient, is a poor comparison. Its only form is a deictic point, indicating attention, but this point, unlike similar deictic expressions in head gesture (head points), manual gesture (hand points), or speech (demonstratives), it is more or less always 'turned on,' during conversation.

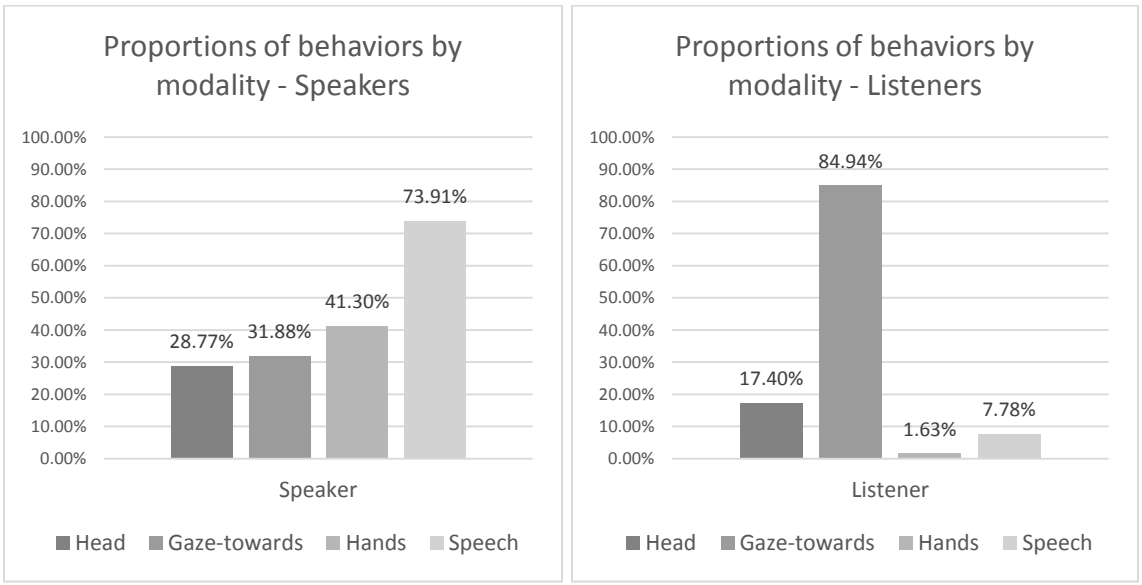
since manual gesture tends to be co-speech gesture[10], and speakers are, of course, speaking more. More surprising is that listener manual gesture is *so* relatively scarce. Speaker manual gesture segments occur around once every six seconds, and these are often complex chains of gestures, while listener manual gesture segments occur less than once a minute. Most interesting, though, is the fact that, without exception, all listener manual gestures occurred alongside speech. One can imagine non-co-speech manual gestures that a listener might make, such as gesturing towards a speaker with half-rotated hand as a sort of affirming back-channel, but no such gestures occurred. It may also be a function of the story-telling context, and non-co-speech manual gestures may occur in more conversational contexts, where they play a role in the negotiation of turn-taking, but there is no a priori reason why manual gesture should be confined to co-speech gesture in a story-telling context. It is an even more striking finding when compared to the highly frequent occurrences of non-co-speech head gestures in listeners. Head gestures often do accompany spoken back-channels, but they are as likely to occur on their own as not.

Distribution of gaze will be more fully covered in Chapter 4, Section 2, but I will note the large, significant difference in the average durations of gaze towards the interlocutor, across roles ($t(1406) = 18.13$, $p < .001$). Listeners spent a great deal more time looking at speakers than vice versa. Correspondingly, they also shifted their gaze less often, once every thirteen seconds, compared to once every six, for speakers. Gaze was not adjusted for within-role exclusions (i.e. excluding behaviors that are produced alongside non-congruent role speech – speakers back-channeling or listener taking speech turns), as it

[10] Manual gesture was not analyzed for functional categories, such as beats, representational gestures, or mimetic gestures, but as a qualitative analysis, it seemed that the majority of speaker manual gesture in this corpus was co-speech, although there was use of pantomime.

does not map as neatly to a speech segment as manual or head gesture, but we will see in Chapter 7, Section 5 that listeners were more likely to shift their gaze away from speakers when they took a speech turn.

Figure 1. Proportions of behaviors by modality: Speakers and Listeners



In addition to rates of behaviors, we can look at how much time participants spent engaged in each modality, as a proportion of the total corpus (Figure 1). These are data we expect to be strongly influenced by the narrative nature of the communicative context.

For example, in the story-telling context, we see that speakers spend around 74% of the time speaking, while listeners are only speaking 8% of the time, as we would expect when one person is the dedicated story-teller. Listeners do produce both back-channels and full speech turns, and listeners' back-channel speech segments tend to be relatively short, compared to their speech turns. There are a number of listener speech turns that are omitted from Figure 1, which make up only 22% of the total listener speech segments, but 43% of the total duration of listener speech segments.

Another major difference across roles is the amount of time interlocutors spend looking at each other. Listeners spend an average of 85% of the time looking at the speaker's face, while speakers only look at the listener's face around 32% of the time, a statistically significant difference ($z = 46.07$, $p < .001$). The proportion of speaker gaze is very similar to the 31% mean gaze time for listeners found in Bavelas et al. 2002. One important thing to note about this difference is that the relatively lower proportion of gaze from the speaker means that, when the speaker *does* look at the listener after a period of gazing away, their gaze is a more salient cue than the listener's gaze and, as we will see in Chapter 7, Section 4, is more likely to lead to a visible response. When listeners look away, it is often alongside a speech turn, so the pattern of looking away while taking a speech turn seems to hold across roles in this speech context. It has been found that gaze patterns between interlocutors are especially culturally dependent (e.g. Levinson & Brown 2016), so researchers with access to video recordings of story-telling from other cultures may find very interesting and very different results.

The differences in proportion of time for manual gesture are even more striking than those of the rates. We see that speakers spend around 41% of their time engaged in manual gesture, while listeners' manual gesture is negligible, at less than 2% ($z = -50.47$, $p < .001$). We will later see that the majority of speakers' manual gesture overlaps with their speech, which means that story-tellers are producing co-speech gesture during more than 50% of their speech, compared to 21% for listeners. Compared with the rates of listener behavior in the modalities of speech, head gesture, and gaze, it seems that manual gesture is not as integral a component of the story-listener role as it is for the story-teller.

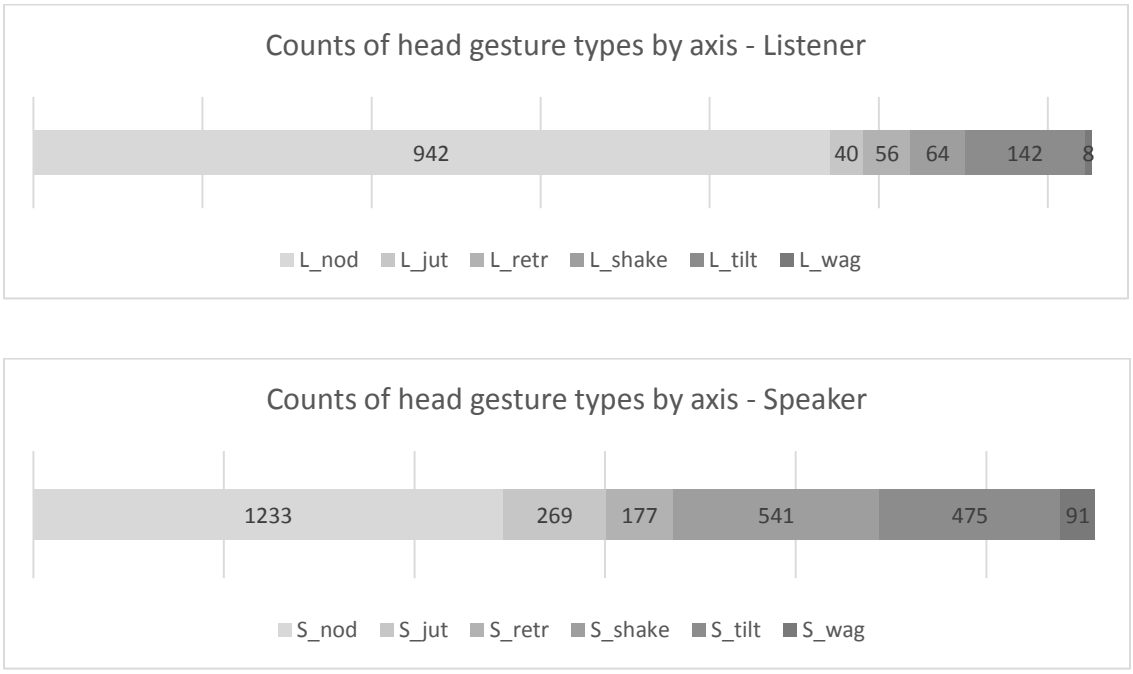
We saw earlier that speakers used head gestures at a more frequent rate, but that listeners used longer head gestures. Proportionally, speakers spend more time engaged in head gesture, although here the relative difference is not nearly as pronounced as in other modalities. Of all the modalities that involve *doing* an action, rather than *holding* a state (that is, all except gaze), head gesture is the least frequent for speakers, and the most frequent for listeners. Given the greater overall activity of speakers, this modality is where the two roles look, at a broad glance, most similar to each other. But many of the really interesting aspects of head gesture involves the subcategories, which we turn to now.

4. Rates of Head Gesture by Axis and Cyclicity

Head gestures are first subdivided by axis of motion, with the categories of nod, shake, tilt, wag, jut, and retraction. Turns of the head are excluded, on the basis that they predominantly co-occur with gaze shift, which is judged to be the more salient behavior. Composite head gestures, on which there is motion on more than one axis at the same time, are also excluded from these figures[11], as are any type for which there were fewer than five instances.

[11] For listeners, composite gestures make up 8% of head gestures; for speakers, composite gestures make up 18% of head gestures.

Figure 2. Counts of head gesture type by axis

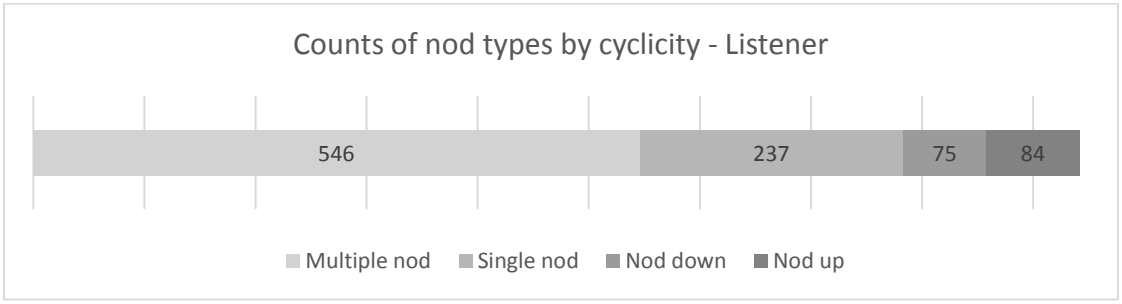


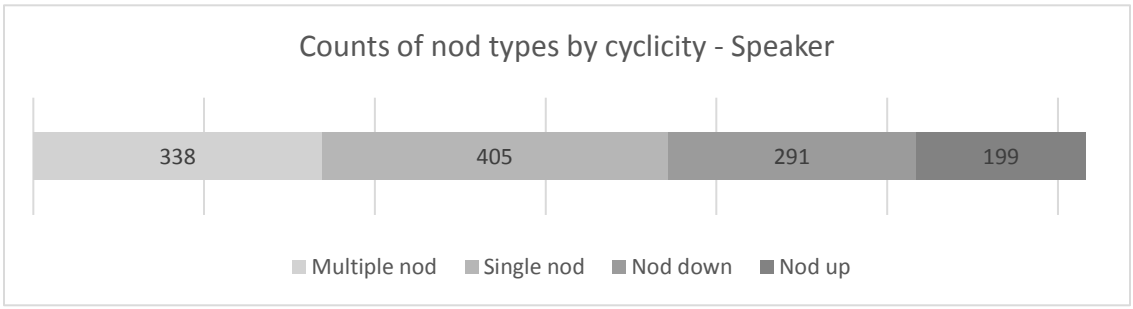
Comparing distributions of types of head gesture across roles (Figure 2), the first thing to note is that listeners have a strong preference for nods, which account for 75% of their total head gestures. This preference is followed by tilts, shakes, retractions, and juts, with only a very small number of wags. Speakers also have a preference for nods, but these make up only 45% of their head gestures. The distributions differ significantly across role (χ^2 (5, N=3456) = 367.10, $p < .001$). Overall, speakers' head gestures are more evenly distributed across the different subcategories. The pattern of listeners preferring a single behavioral category and speakers having more evenly distributed category is quite interesting – it is seen at almost every level of analysis. If we assume that different formal categories of head gesture correspond to different functions, then this difference in the variability across roles makes sense. Speakers have a greater variety of functions to communicate. However, I don't wish to suggest that listener head movements are monolithic, serving only a single purpose. Even a coding scheme as simple as the one

used here can capture large differences in frequencies of types, which suggest differences in functional types as well.

4.1 Nods

Figure 3. Counts of nod types by cyclicity





Looking at subtypes of nods, broken down by cyclicity (Figure 3), listeners exhibit the same pattern seen in Figure 2, with multiple nods accounting for nearly 60% of their nod types. And, also following this pattern, speakers' nod types are evenly distributed across the four categories, with a significant difference in distribution across roles: χ^2 (3, N=2175) = 232.33, $p < .001$.

4.2 Shakes

Figure 4. Counts of shake types by cyclicity

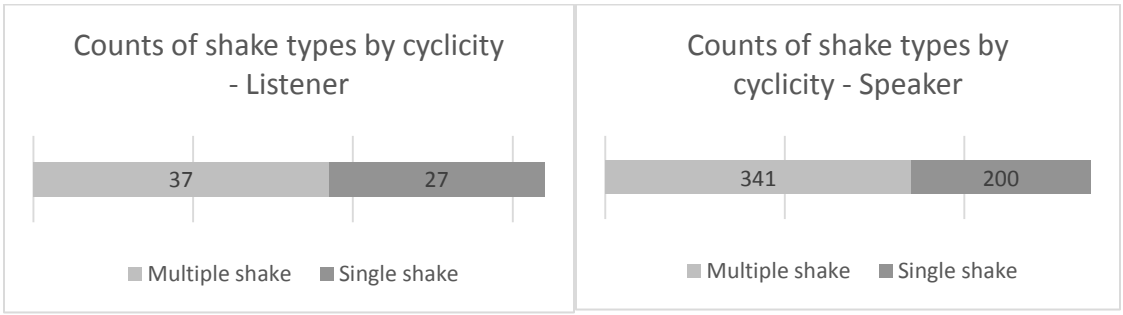


Figure 4 show the distributions of shake types across roles. This is the only behavior which does not show a significant difference across roles, χ^2 (1, N=605) = 0.46, p = 0.42, although both listeners and speakers both show a preference for multiple shakes.

4.3 Juts and retractions

Figure 5. Counts of jut and retraction types by cyclicity

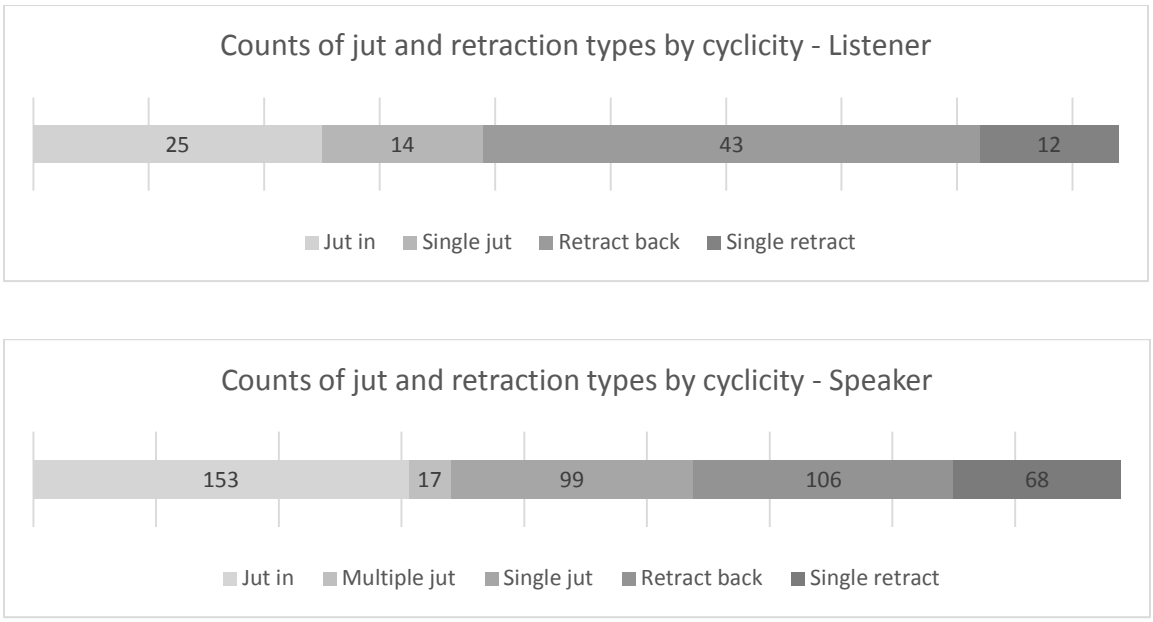


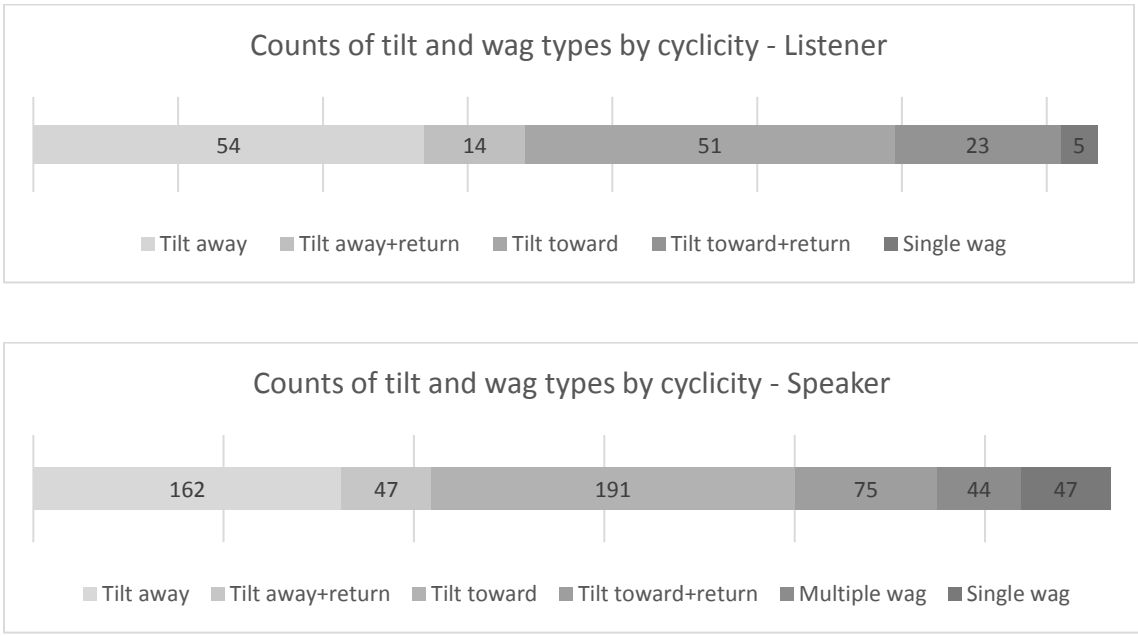
Figure 5 show the distributions of subtypes of juts and retractions across roles, with the distributions being significantly different across roles, χ^2 (3, N=520) = 16.63, p < .001. Multiple retractions are excluded from both figures because there were fewer than 5

tokens, as were multiple juts from Figure 5. Juts and retractions have been grouped together in these figures to show differences in preference for speakers and listeners. Retractions make up nearly 60% of listeners' gestures on this axis, while juts make up over 60% of speakers' gestures. Again, listeners have stronger preferences than speakers for a single category. For retractions, they prefer retracts-back over the single retraction, also prefer the jut-in over the single jut. Speakers' retractions are more evenly distributed across retracts-back and single retractions, but show a slight preference for juts-in over single juts (multiple juts are dispreferred).

Juts and retractions both occur on the same line of displacement, the z-axis. This seems like it would be an especially salient axis in communication, as it is the axis connecting the two participants. It may be that juts and retractions iconically map to the flow of information. According to this model, juts would correspond to the conveying of information, in the same sense that information metaphorically moves towards the listener, and retractions would correspond to the receiving of information. This might help explain the fact that we see a significantly greater proportion of juts from speakers, and of retractions from listeners (χ^2 (1, N=537) = 10.95, $p < .001$), given their respective roles in the transfer of information in a story-telling context. There is certainly more to the story, though. Retractions, like nods-up, can also correspond to surprise or other affective responses.

4.4 Tilts and wags

Figure 6. Counts of tilt and wag types by cyclicity

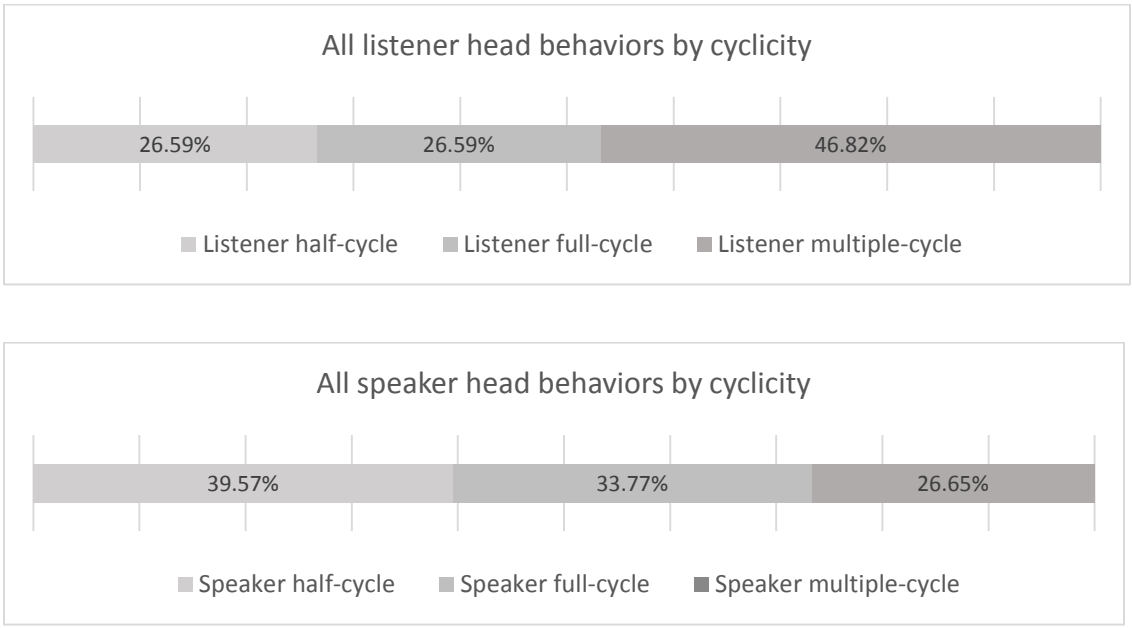


Figures 6 show the distributions of types of head motion around the z-axis. This is the category with the least amount of difference across roles, with speaker and listener usage of different types being more or less equally distributed, with no significant differences across role. In both groups, there was a preference for half-rotations, though no preference for tilting towards or away. It seems possible that this distinction is meaningful, and we will look at how half-cycle head gestures pattern together in Chapter 5, Section 3.

4.5 All Head Behaviors

The other primary dimension in the head gesture coding scheme is cyclicity. If we collapse across axis of motion, we can group all half-cycles, full cycles, and multiple cycles together and look at these distributions across roles.

Figure 7. All listener head behaviors by cyclicity



In Figure 7, we see that listeners and speakers differ significantly in their preference for different cycles of head gesture, χ^2 (2, N=4051) = 162.27, $p < .001$. Listeners, whose preferred head gesture is the multiple nod, tend more towards multiple cycles, and are evenly split between half-cycles and full cycles. Speakers, again more evenly distributed across categories, have a stronger preference for half-cycles, with multiple cycles being the least commonly used. We can speculate about the causes for these differences. Listeners may prefer longer, cyclic gestures because these are easier for speakers to detect, given speakers' sporadic glances towards the listener. Half-cycles (which will often be referred to as 'repositioning' gestures), such as nods up or tilts away, often signal a shift in perspective or thought. Speakers may prefer these gestures because, as narrators, they are rapidly shifting from one perspective to another over the course of the narration.

5. Durations of Subtypes of Head Gesture

We have looked at rates and durations of behaviors at the broad level of modality, and we examined rates of behavior at the finer-grained subtypes of head gesture. Table 4 shows mean durations of each subtype of head gesture by role, with frequencies and rates included for reference. This table includes composite behaviors, each dimension of a composite gesture is counted separately (as unique composite gestures are too numerous to list). The M diff. column shows the value of the listener mean minus the speaker mean, with high positive values indicating longer average durations for listeners, high negative values indicating longer average durations for speakers, and low values indicating similar averages across role.

Table 4. Summary statistics of subtypes of head gesture by role

	Listener				Speaker				
	Coun	Rate	M	SD	Coun	Rate/	M	SD	M
Multiple nod	546	5.72	1123.9	782.24	338	3.54	1116.9	683.3	6.97
Single nod	237	2.48	608.59	366.15	405	4.24	643.05	364.5	-34.45
Nod-down	75	0.79	529.80	502.28	291	3.05	490.46	341.0	39.33
Nod-up	84	0.88	441.65	256.16	199	2.08	511.23	318.5	-69.57
Multiple	37	0.39	1346.8	1035.5	341	3.57	1052.2	617.4	294.60
Single shake	27	0.28	513.59	206.26	200	2.09	537.18	281.7	-23.58
Jut-in	25	0.26	460.00	211.68	153	1.60	481.46	314.6	-21.45
Single jut	14	0.15	1104.7	697.25	99	1.04	811.45	399.9	293.33
Multiple jut	1	0.01	700.00		17	0.18	1223.5	624.9	-
Retract-back	43	0.45	541.88	294.81	106	1.11	520.14	280.3	21.74
Single retr.	12	0.13	969.42	397.51	68	0.71	803.96	409.3	165.46
Multiple retr.	1	0.01	933.00		3	0.03	1122.3	38.68	-
Tilt away	54	0.57	566.07	308.61	162	1.70	505.33	282.9	60.74
Tilt away-ret.	14	0.15	1035.7	471.74	47	0.49	740.45	318.5	295.26
Tilt towards	51	0.53	611.75	346.88	191	2.00	499.15	310.6	112.59
Tilt towards-	23	0.24	717.43	291.95	75	0.79	701.36	314.2	16.07
Multiple wag	3	0.03	1000.0	260.36	44	0.46	1265.8	541.8	-
Single wag	5	0.05	433.00	100.00	47	0.49	602.85	276.7	-169.8

Listeners have significantly longer average durations for single juts (M=1104.79, SD=697.25) than speakers do (M=811.45, SD=399.94), $t(111) = 2.31$, $p = 0.02$, around 20-30% longer. Listeners also tend to have longer multiple shakes (M=1346.89, SD=1035.56) than speakers (M=1052.29, SD=617.44), $t(376) = 2.54$, $p = 0.01$, which can extend, like a multiple nod, over multiple seconds of story-telling. Listeners also have significantly longer tilts towards (M=611.75, SD=346.88) than speakers (M=499.15, SD=310.60), $t(240) = 2.24$, $p = 0.03$, and also tilts-away-and-return (M=1035.71, SD=471.74) than speakers (M=740.45, SD=318.59), $t(59) = 2.71$, $p = 0.01$, but do not differ much in the average duration of tilts-towards-and-return. These two kinds of tilts are similar in that both finish with motion towards the interlocutor. But perhaps the most striking finding is the similarity in duration across roles

This is particularly noticeable in nods. While there are some small differences in average durations of nods (speakers' single nods and nods-up tend to be slightly longer, and their nods down tend to be slightly shorter, but not significantly so), speakers and listeners do not differ substantially in their average duration of multiple nods. These are the most common type of head gesture for listeners, and the second most common for speakers, and also one of the longest head gestures (1124ms for listeners and 1117ms for speakers), where we might expect to see greater differences. Despite the nearly identical length, an examination of the ways that speakers and listeners are using this gesture suggest that it is being used for different functions – for listeners it tends to function as a continuer and acknowledgement, and for speakers it tends to function as a way of emphasizing or marking the speech content they are delivering. One possible reason for the similarity in length is that they both correlate with some other behavior of the same length. Another possible theory is that these are not completely different functions, that a listener multiple nod is simply one person emphasizing or marking the information of another person's speech. As they process the information they are receiving, they are taking on the mental model of their interlocutor, and emphasizing it as it is processed. One can easily imagine the communicative scenario in which one person is speaking and their listener guesses what they are going to say, and begins nodding excitedly, often in concert with the speaker. Or this may have been the origin of the listener nod, which eventually became conventionalized from an informative indicator of processing to an (arguably more important) communicative indicator of attention. It is certainly the case that one learned convention in many cultures is to nod even when we are not paying attention.

6. Duration and Rates of Speech by Subtypes

6.1 Speech Turns

We have seen the distributions across roles of the broad category of speech. We turn now to two sets of subcategorizations of speech segments. First, we look at subtypes of speech turns examining patterns within these subtypes and with spoken back-channels. In these analyses, we will look at *all* speech segments, not only those that are congruent with each participant's role. That is, we will see turns and back-channels from both story-tellers and story-listeners.

Table 5. Summary statistics of subtypes of speech by role

	Listener				**Speaker**				
	Count	**Rate/m**	**M**	**SD**	**Count**	**Rate/m**	**M**	**SD**	**M Diff.**
Backchannel	597	6.79	582.37	414.04	102	1.16	521.22	350.91	61.15
Declarative	99	1.13	2005.38	999.46	1050	11.93	2778.86	1766.16	-773.48
Interrogative	82	0.93	1435.37	810.93	217	2.47	2798.31	1617.27	-1362.94
Filler	5	0.06	266.60	99.83	263	2.99	586.17	335.64	-319.57
Incomplete	18	0.20	936.94	553.68	349	3.97	1487.97	1069.38	-551.03

Table 5 shows the distributions of speech turn subtypes for speakers and listeners. Speech turns are broken into four subtypes: declaratives (which complete a predication and do not have question intonation or syntax), interrogatives (which complete a predication and *do* have question intonation or syntax), fillers (which consist of filled pauses and discourse connectors like *so* or *then*), and incompletes (which begin or continue a predication, but do not complete it). Non-turn behavior is grouped together as spoken back-channels and laughs (see Table 6 below).

Comparing rates of speech turn behaviors across roles, an obvious finding is that speakers are taking more speech turns of all subtypes. They use declaratives ten times more frequently than listeners, and interrogatives two and a half times more frequently[12]. Speakers' declaratives are four times more frequent than their fillers, and three times more frequent than their incomplete speech segments. Listeners do not use fillers or incomplete segments frequently enough to explore any patterns. This is not surprising, as these subtypes are features of longer utterances. However, it is clear that listeners are indeed participating in the stories in the form of full speech turns.

Comparing rates of back-channels across roles, we see that this is the primary speech behavior for listeners, but that speakers also occasionally produce back-channels (mostly in response to declarative or interrogative speech from listeners).

In terms of the responsiveness between speech turns and back-channels, is there a difference across roles? Looking at the ratios of turns to back-channels across roles, we see that speakers back-channel at a rate of 1.16 / minute and listeners back-channel at a rate of 6.79 / minute, and that speakers produce declaratives and interrogatives[13] at a rate of 14.4 / minute while listeners produce them at a rate of 2.06 / minute. The ratio of speaker back-channel rate to listener speech turn rate is 56% and the ratio of listener back-channel to speaker speech turn rate is 47%, suggesting that listeners are less verbally responsive to speakers than vice versa.

[12] Listeners' interrogatives are much more likely to have question syntax than question intonation, and speakers show the reverse pattern.

[13] These two speech turn types are much more likely to elicit back-channels, as we will see in Chapter 7, Section 3.

Looking at the differences in average duration, it is clear that speakers' average durations are much longer for all speech turn subtypes. Even when a listener takes a speech turn in a story-telling context, it is relatively reduced compared to the story-teller. The only case where this is not true is with back-channels. Listeners' back-channels are slightly shorter (M=582.37, SD=414.04) than speakers (M=521.22, SD=350.91), but not significantly so, $t(697) = 1.41$, $p = 0.16$. This may be due to the fact that the set of constructions and lexical items involved in spoken back-channels is quite constrained, and similarly constrained for both roles.

6.2 Back-channels

We can also look at the distribution of subtypes of spoken back-channels across roles (Table 6). As in the previous analysis, we are including all speech behaviors, including non-congruent ones. We have also excluded two categories of back-channels (clarification questions and brief question units) for having too few tokens in both roles.

Table 6. Summary statistics of back-channel subtypes by role

	Listener				**Speaker**			
	Count	**Rate/m**	**M**	**SD**	**Count**	**Rate/m**	**M**	**SD**
Acknowledgement	204	2.32	438.75	292.64	12	0.14	775.00	494.31
Affirmation	41	0.47	546.34	398.43	75	0.85	499.51	358.77
Assessment	222	2.52	797.74	463.54	3	0.03	989.00	428.68
Collaborative finish	21	0.24	911.10	334.48	3	0.03	655.67	138.91
Continuer	95	1.08	368.44	147.73	0	0.00	--	--
News-marker	20	0.23	578.40	316.00	1	0.01	500.00	--
Laugh	125	1.38	635.83	478.25	37	0.27	823.58	413.61

Listeners exhibit a great variety in the kinds of back-channels they use. Acknowledgments and assessments are the most frequently used, at two and half per minute, followed by continuers, which are used around once per minute. Speakers engage in only a small number of back-channel types, primarily limited to affirmations (responding to listener interrogatives) and acknowledgments. Comparison of durations across roles shows no significant difference between any back-channel types except acknowledgements, where speakers (M=775, SD=494.31) are significantly longer than listeners (M=438.75, SD=292.64), $t(214) = 3.70$, $p < .001$.

We see that listeners engage in stand-alone laughs as an utterance type much more frequently than speakers[14], and many of these laughs appear to fill the function of a back-channel response (some researchers have classified laughter as a form of back-channel, such as Maynard (1997)). Speakers tend to laugh, on average (M=823.58, SD=413.61), longer than listeners (M=635.83, SD=478.25), $t(160) = 1.79$, $p = 0.75$, although this is only marginally significant. Laughs are another example of a category that is often seen as having a monolithic function, but which has enormous and unexplained variety, both in function and form.

This concludes the summary statistics of the modalities in this data set. The remaining chapters will deal with timing patterns of these modalities. Chapter 4 will look at the timing of individual modalities within speakers and listeners, while chapters 5 through 7 will look at the timing interactions between modalities, both within role and across role.

[14] Speakers' laughs are underreported in this data set, as they frequently are part of longer utterances, often occurring simultaneously with a word.

CHAPTER IV: WITHIN-ROLE / WITHIN-MODALITY

1. Introduction

This chapter looks at the timing relations of behaviors that are both within the same modality and within the role of speaker or listener, such as the timing of listener head nods with other listener head nods. The methods used in this chapter differ slightly from those used in Chapters 5 through 7. Those chapters look at the timing relations and overlaps that are between behavior boundaries that are across role or across modality, which are not feasible for within-role, within-modality analyses. Onsets and offsets of the same behaviors may occur near each other, and we will examine these patterns, but the bulk of the temporal patterns we can look at are across-category, for purely mathematical reasons.

This is also a style of analysis that has already been well-documented. The patterns of speech within an individual make up the bulk of linguistic research. The other three modalities do not have the same breadth of scholarly research as speech, but their individual characteristics have been documented in many ways. And this is sensible: it is a good idea to begin by analyzing patterns of one behavior in isolation from other patterns before looking at their interactions.

Section 2 will explore the patterns of timing between offsets and onsets of a single behavior within a single role. These lag times differ considerably across modalities, and across individuals. Section 3 will examine n-gram sequences of subtypes of behaviors within a single modality, looking at variation in sequential patterns for each role. Section 4 looks at the timing relations of behaviors as relate to the particular communicative context we are analyzing (story-telling). The likelihood of each behavior occurring is not

necessarily constant at every given moment during a story, but is tied in some way to the nature of the story-telling event. These different patterns can be seen by looking at frequency distributions of behaviors over the time-course of the story-telling. Section 5 summarizes the findings and suggests some hypotheses derived from the data that might warrant further qualitative analysis.

2. Timing Patterns of Lag Between Behaviors

We have already looked at the corollary analysis to this in Chapter 3, where we saw the durations of behaviors in each modality, i.e., looking at the distance between onsets and offsets of behaviors. Here we will look at the distance between offsets and the following onsets of specific behaviors, the pause or lag times. This analysis differs substantially from the previous in that, while we have a number of categorical distinctions within different modalities, we do not have any categorical distinctions of pauses. For this reason, I will look only at the lag times between the four broad categories of modality rather than between subtypes. This analysis will also be including non-congruent behaviors (e.g. listeners taking speech turns), which will make certain kinds of interpretations difficult. Doing so will obscure temporal tendencies that are specific to a particular role, but omitting the non-congruent behaviors would obscure temporal tendencies of the modalities overall. Since it is not clear to what extent these patterns are dependent on role, we have opted to include everything. Additionally, some behaviors (particularly heads and hands) occur in chains with little to no lag time between offsets and following onsets, and these pauses have also been omitted. Finally, the beginning points and end points of stories were counted as boundaries (so the first listener head lag

time in a story might be from the beginning of the story to the onset of the first listener head segment)[15].

Table 7 below shows the summary statistics of lag times between behavior offsets and the following onset for that behavior. Many of these distributions are heavily negatively skewed, much like the distributions of behavior durations (as well as many other distributions in this data set). Outliers greater than 3 standard deviations from the mean of the log transform of each behavior category were omitted[16], all being greater than the mean.

Table 7. Summary statistics of lag times between behaviors by modality and role (ms.) (non-congruent behaviors included)

		Count	M.	SD	Min	Median	Max
Listener	**Head**	1037	3898	3601	300	2700	19,000
Listener	**Gaze-away**	438	1633	1152	500	1267	7467
Listener	**Hands**	107	49,486	53,350	900	32,033	212,900
Listener	**Speech**	748	6030	5631	233	4233	32,133
Speaker	**Head**	1723	1960	1788	300	1333	9567
Speaker	**Gaze-away**	988	3364	2871	500	2433	15,967
Speaker	**Hands**	795	3415	3998	500	1800	23,600
Speaker	**Speech**	1841	658	527	200	500	3600

One pattern we see (comparing Table 7 with Table 3) is that, for most modalities, the duration of the pause between behaviors and the duration of the behaviors themselves are inversely related for each role – listeners tend to produce shorter segments and have

[15] So several modalities will have different frequencies of lag times than behavior durations in Table 3.
[16] The proportion of outliers were less than 3% for each role and modality combination.

longer pauses between them, while speakers tend to produce longer segments and have shorter pauses between them. The greatest difference is seen in lag times between manual gesture segments across roles, with listeners' lag times (*M*=49,486ms.) being on average 15 times longer than speakers' (*M*=3415ms.), more evidence of how differently this function is used across roles.

There is also a large difference in the average lag time between speech segments, with listeners (*M*=6030ms.) lagging nearly ten times as long as speakers (*M*=658ms.) between speech segments (back-channels or turns), and with a large amount of variance for both roles. Across roles there is a much smaller difference between head gesture lag times (average lag times for listener head gestures (*M*=3898ms.) are around twice as long as for speakers' (*M*=1960ms.)), when compared to the differences between lag times for hands and speech.

The greater similarity across roles for the timing of heads is similar to what we saw in Table 3, where durations of head gestures across roles showed the least difference of all four modalities. There it was suggested that the similarity of durations might have been the result of physical or functional constraints of head motion: the number of possible forms, and therefore possibly functions, is relatively reduced compared to speech and manual gesture. But neither of these explanations would account for the greater similarity in lag time across roles. It may be that listener head gestures are slightly more frequent than listener speech because the 'windows of opportunity' (the periods of time in which the behavior can be detected, visibly or audibly, and in which it will not impede communication) are slightly more frequent for visible behaviors (speakers look at listeners around 32% of the time) than for audible behaviors (speakers are not speaking

around 26% of the time). Or it may be that that they are more frequent because they are serving more functions than speech, and some of these functions may not be communicative: there may be cases of nodding that have more to do with processing meaning (something done more or less continuously during story-listening) than communicating attention (although it's possible the same may be true for some spoken back-channels). Or it may be that there is some sort of interactive relationship between speaker and listener head behaviors that is partially independent of what is happening in the speech stream. Or it may be some combination of these.

The modality that stands out as the most different from the durations in Table 3 (that is, it is not merely negatively correlated) is eye gaze. The lag time between the offset of gaze and the next onset effectively measures the periods of time in which participants are looking away from each other. Here, as with lag time between head gestures, speakers and listeners are much more similar to each other, although the difference is in the opposite direction. Listeners' gaze-aways are on average half the length of speakers' gaze-aways. This is in stark contrast to the durations of gaze-towards, where listeners' average gaze-towards (8054ms.) were around five times as long speakers' gaze-towards (1647ms.) (we will see in Chapter 5, Section 2 how much of this time was spent looking at each other). What this looks like over the course of a story is listeners spending most of their time looking at the speaker, and *occasionally* taking brief, one to two second glances away, often when they take a speech turn or produce an assessment back-channel, and speakers spending much less time looking at the listener, but *frequently*

making brief, one to two second glances towards the listener[17]. The fact that listener gaze-away (1632ms) and speaker gaze-towards (1647ms) are of such similar durations is curious, given that they seem to be fulfilling different functions.

3. Sequences of Behaviors ("n-grams")

So far in this chapter, and in Chapter 3, we have mostly been analyzing behavior segments over units of time – how many seconds (or frames) are in a segment or in a lag between segments, and what is the ratio of segments to overall time. But one important aspect of a system of communicative interaction is the sequential patterns of the units themselves. In this section we will consider behaviors as units that occur in sequences, much like n-grams in a text corpus. The n-gram analysis in this chapter, looking at behaviors within the same modality and within role, will deal with bigrams of head and speech behaviors. Chapter 5 will examine sequences of behaviors across modality for the same participant, Chapter 6 will look at sequences within the same modality between speakers and listeners, and Chapter 7 will look at sequences across both modalities and roles. For a description of n-gram analysis, please refer to Chapter 2, Section 6.

Looking at behavior sequences within modality can be done locally or globally, locally in the sense of looking at behaviors that occur with a certain time span, and globally in the sense of looking at sequences across the entire story, irrespective of how near they are to each other. Each of these approaches tells us different information. Within-modality sequences that occur in close proximity to each other can be useful in hypothesizing about production planning, and what kinds of syntagmatic dependencies there are in

[17] Although note the substantial amount of variance. Some speakers, particularly males spent long periods of time looking away, with infrequent gaze-towards, and four listeners (three female) never once looked away over the course of the entire story (these latter were excluded as outliers from Table 7.

different modalities, or what kinds of functions these collocations might have. Looking at sequences at the global level can be useful in hypothesizing about the larger constructional units that make up a narrative or other communicative context. Additionally, both of these kinds of sequences can be compared to across-role sequences, to help disentangle the extent to which patterns are better explained by internal or interactional patterns.

In this chapter, analyses will be limited to local analyses, and to bigrams. 3-grams and 4-grams were examined, but one would need a much larger corpus than this to provide any interesting results that are not already captured by the bigram analysis.

For these sequences, where we will be looking at bigrams of behaviors offsets followed by behavior onsets, in close proximity to each other, some span of time needs to be defined. The size of this window could be motivated by a number of things, like average duration of the behaviors, motoric planning, and the nature of the dataset. Unfortunately, to my knowledge, there are no comparable studies that have ever employed a methodology like this. For these analyses, I have chosen a three-second window, which is narrow enough to be proximate, and broad enough to capture enough tokens to analyze. We will also be able to look at the conditional probabilities of offsets and onsets in each bigram pair, using the symmetric conditional probability of the bigrams to assess their dependency[18].

[18] We do not have odds ratios for local pairs. This would require determining the expected overlap of two behaviors based both on their frequencies relative to the span of the corpus, and relative to the 3-second span. Work on such a method is ongoing.

3.1 Bigrams of Speaker and Listener Head Subtypes

To begin, we look at bigrams of head gestures in speakers. Tables 4.2a/b show the 15 most frequent bigrams of head offsets followed by head onsets, ordered by their symmetric conditional probability.

Table 8. Frequencies and symmetric conditional probabilities of local head bigrams

Speaker head bigrams				**Listener head bigrams**		
Rank	**Offset + Onset**	**Freq.**	**Symmetric CP**	**Onset + Offset**	**Freq.**	**Symmetric CP**
1	nod down + nod up	31	0.035	m. nod + m. nod	140	0.078
2	m. nod + m. nod	40	0.030	nod down + nod up	19	0.077
3	m. shake + m. shake	30	0.023	m. shake + m. shake	5	0.034
4	s. nod + m. nod	31	0.015	s. nod + s. nod	29	0.021
5	s. nod + s. nod	32	0.013	m. nod + s. nod	44	0.019
6	nod down + nod down	23	0.013	s. nod + m. nod	40	0.016
7	s. nod + tilt away	16	0.010	m. nod + nod down	18	0.010
8	m. shake + s. shake	15	0.010	nod up + nod down	5	0.006
9	s. nod + nod down	23	0.009	s. nod + retract back	5	0.005
10	nod up + nod down	16	0.009	s. nod + nod down	8	0.005
11	nod up + m. nod	17	0.009	m. nod + nod up	13	0.005
12	m. nod + nod down	19	0.008	m. nod + retract back	7	0.004
13	m. nod + tilt towards	14	0.007	nod up + s. nod	7	0.003
14	nod down + s. nod	20	0.007	m. nod + tilt away	8	0.003
15	tilt away + s. nod	13	0.007	s. nod + tilt away	5	0.003

One thing that stands out is that some of the strongest dependencies we see in head bigrams are between pairs of the same kinds of behaviors. Bigram pairs of multiple nods, multiple shakes, and single nods are among the five highest ranked for symmetric conditional probability for both speakers and listener, and pairs of nods-down are ranked

6 for speakers. Whether accompanying speech (for speakers, primarily) or responding to speech (for listeners, primarily), participants have a tendency to repeat these same head behaviors in succession. For multiple nods and shakes, the magnitude of the gesture most often starts large and diminishes with successive cycles until it becomes nothing, so in these bigrams the second multiple nod or shake may be a continuation of the same communicative function, but restarted for some reason. This same principle might be at work with single nods and nods down.

Another point to be made is that there are a number of subtypes of gesture that are absent from the more strongly dependent bigrams. For speakers, there are no bigrams with wags, juts, or retractions, and for listeners there are none with wags or juts. Half-cycle tilts do seem to occur in sequence with other head gestures. For listeners, several half-cycle gestures tend to follow nods – nods-down, nods-up, tilts away, and retractions back follow different nod types in nine of the fifteen highest ranked listener bigrams. Half-cycle head gestures, which reposition the head, can be interpreted as indicating a shift of perspective. In the case of these bigrams, it may be that the construction of the nod and the repositioning gesture indicate something like 'I have heard what you have said, and now I shift my head to take in what you will say.' Or it may be that this kind of bigram occurs when the speaker is making a narrative shift – this is a testable hypothesis (see section 5.2).

3.2 Bigrams of Speaker and Listener Speech Subtypes

We now look at bigram of speech types. The left columns in Table 9 collapse back-channel types into a single type (mostly affirmations), while the right columns in Table 9 include all speech types, including turn types and back-channel types.

Table 9. Local frequencies and symmetric conditional probabilities of speech segment bigrams

Speaker speech bigrams				Listener speech bigrams		
Rank	**Offset + onset**	**Freq.**	**Symm. CP**	**Offset + onset**	**Freq.**	**Symm. CP**
1	declarative + declarative	549	0.273	incomplete + declarative	11	0.077
2	incomplete + declarative	203	0.112	declarative + declarative	19	0.040
3	declarative + filler	165	0.099	declarative + incomplete	7	0.034
4	declarative + incomplete	159	0.069	interrogative + interrogative	12	0.026
5	filler + declarative	133	0.064	assessment + assessment	34	0.025
6	back-channel + back-channel	23	0.052	acknowledgment + acknowledgment	28	0.019
7	incomplete + interrogative	58	0.045	laugh + laugh	13	0.013
8	filler + incomplete	58	0.037	assessment + interrogative	13	0.011
9	interrogative + declarative	90	0.035	interrogative + assessment	13	0.010
10	interrogative + interrogative	40	0.034	declarative + assessment	14	0.010
11	declarative + interrogative	83	0.030	affirmation + declarative	6	0.009
12	incomplete + incomplete	59	0.029	laugh + declarative	10	0.008
13	interrogative + incomplete	45	0.027	declarative + affirmation	5	0.007
14	declarative + back-channel	42	0.017	acknowledgment + laugh	12	0.007
15	back-channel + declarative	41	0.016	assessment + declarative	11	0.006
16	interrogative + filler	30	0.016	coll. finish + assessment	5	0.006
17	filler + filler	33	0.016	interrogative + acknowledgment	9	0.005
18	back-channel + laugh	6	0.012	continuer + acknowledgment	10	0.005
19	filler + interrogative	26	0.012	laugh + acknowledgment	11	0.005
20	laugh + back-channel	5	0.007	acknowledgment + assessment	14	0.004

Within the 3-second window, we see stronger dependencies with turn types than back-channels, because back-channels are usually produced more than three seconds apart. For both speakers and listeners, the strongest bigram dependencies we see are between declarative pairs, and between declaratives and their grammatical dependents, fillers and incompletes. Bigram pairs of interrogatives are also fairly highly ranked.

Pairs of back-channels are fairly dependent for speakers (e.g. affirming what a listener is asking during the speech, and then again after it: e.g., *Uh huh.*... *Yeah.*). For listeners, pairs of assessments, acknowledgments, and laughs are among the higher ranked bigrams. As with head gestures, it seems that the strongest dependencies in speech are sequences of the same kind of behavior. These repeated back-channel bigram pairs could be analyzed to see whether they are responding to the same topic in the interlocutor's speech, and how the speech context that precedes the first back-channel differs from the first.

4. Frequency Distribution of Behaviors over the Course of the Story

Communication between people happens in a wide variety of conventionalized communicative contexts, from talking about other people, to arguing about ideas, to relating personal anecdotes. These contexts can differ widely in terms of how interactive they are, the nature of the goals of the participants, the kinds of social pragmatics involved, or the nature of the semantic information being relayed. These differences lead to differences in the structure of the discourse. The structure of story-telling is perhaps the most well-analyzed, typically being comprised of five sequential components: exposition, rising action, a climax, falling action, and resolution. In the exposition, the scene is set, and characters, relationships, and setting are introduced. In the rising action,

conflicts and complications are developed, setting the stage for the climax, which is the part of the story that is the peak of tension and emotional intensity. This is followed by the falling action, where the results of the climax are described, and the resolution, where the threads of the story are tied off.

This structure is drawn from analysis of literary texts, but spoken stories follow the same structure (Tannen 1982). Two major differences between a spoken story and a written or performed story are that, in a spoken story, the story-teller can gauge their audience's response and adjust the telling accordingly, and that the audience can actively participate to make comments or ask questions.

In the forty stories analyzed here, speakers begin by setting the scene, in terms of the characters and location involved, and sometimes in terms of what kind of story the listener should expect[19]. An example is given below.

Example 4.1.

A: *So I have three best friends that I've been, we've been best friends since I, since we were in preschool.*

This introduction is followed by describing the series of events that generated the near death (or embarrassing) climax, which was followed by an explanation of the resulting events (typically an explanation of how they didn't actually die). A close analysis of the narrative structure was not done, but in a subjective analysis of the stories, the climax tended to occur somewhere between 60 and 90% of the way through the story, depending on how much falling action and resolution was involved, and particularly on how many

[19] 34 out of the 40 stories started with the word *so* (or some variant, such as *okay, so* or *all right, so*).

questions the listener asked (the majority of stories ended with listeners asking questions to clarify parts of the story that were unclear and possibly to show interest).

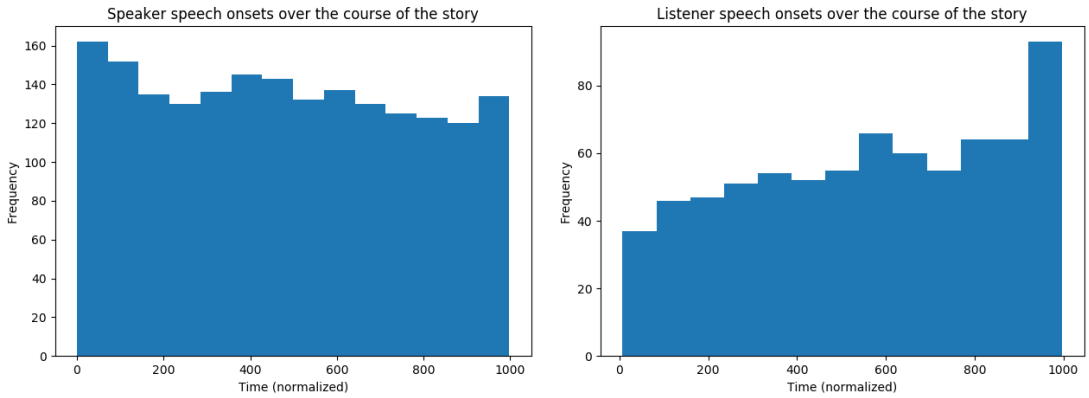
Given what we know of the narrative structure (and, by extrapolation, what we know of the socio-pragmatic goals that are likely to be involved at different parts of the narration) we can look at the distribution of behaviors over the course of the story and see whether these two sets of structures pattern together. The focus of this analysis will be on head and speech behavior, as these are the modalities that are most finely coded here, but this kind of analysis could easily be extended to finer categories of manual gesture or facial expressions.

The following plots, called *story histograms*, show the frequency distribution of different behaviors over the course of the story. Since the stories are of quite different lengths, they have all been normalized to the same length, 1000 frames. This means that some stories are quite compressed, so these plots will not be helpful for interpreting patterns of behavior that rely on absolute timing, only for patterns that rely on timing relative to the story as a whole.

4.1 Speech Behaviors

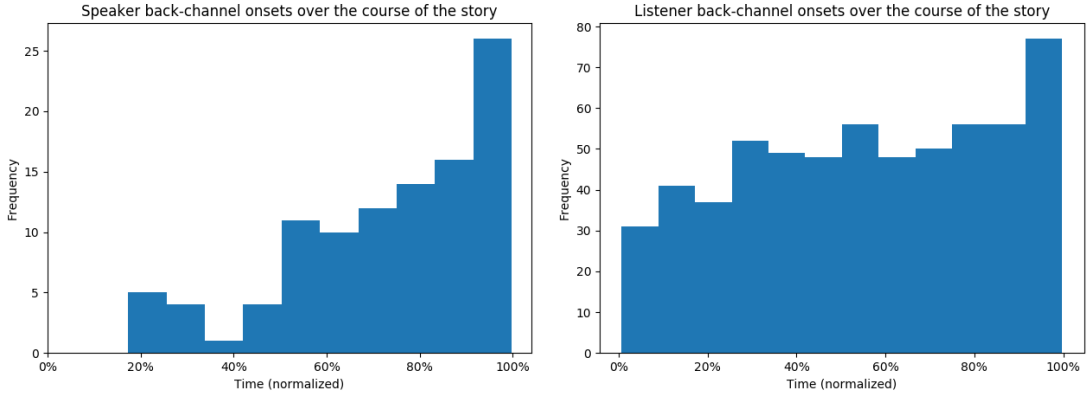
First we'll look at speech behaviors overall, including non-congruent behaviors (listener speech turns and speaker back-channels), then we'll break these down by congruency and category.

Figure 8. Story histogram – Speaker and Listener speech onsets



Looking at all speech onsets across roles, we see that listeners gradually increase their rate of speech segments over the course of the story, with a dramatic rise at the very end of the story. Speakers, on the other hand, are fairly uniform in their frequency across the entire story, with a slight peak at the beginning of the story, since all stories start at the beginning. To look more closely at speech behavior, we'll first look at the distributions of speech turns. Throughout these figures, there will be a number of peaks during the final bin of the histogram. In these stories, this 'wrapping up' portion of the story tended to involve more turn-taking, with follow-up questions and laughter.

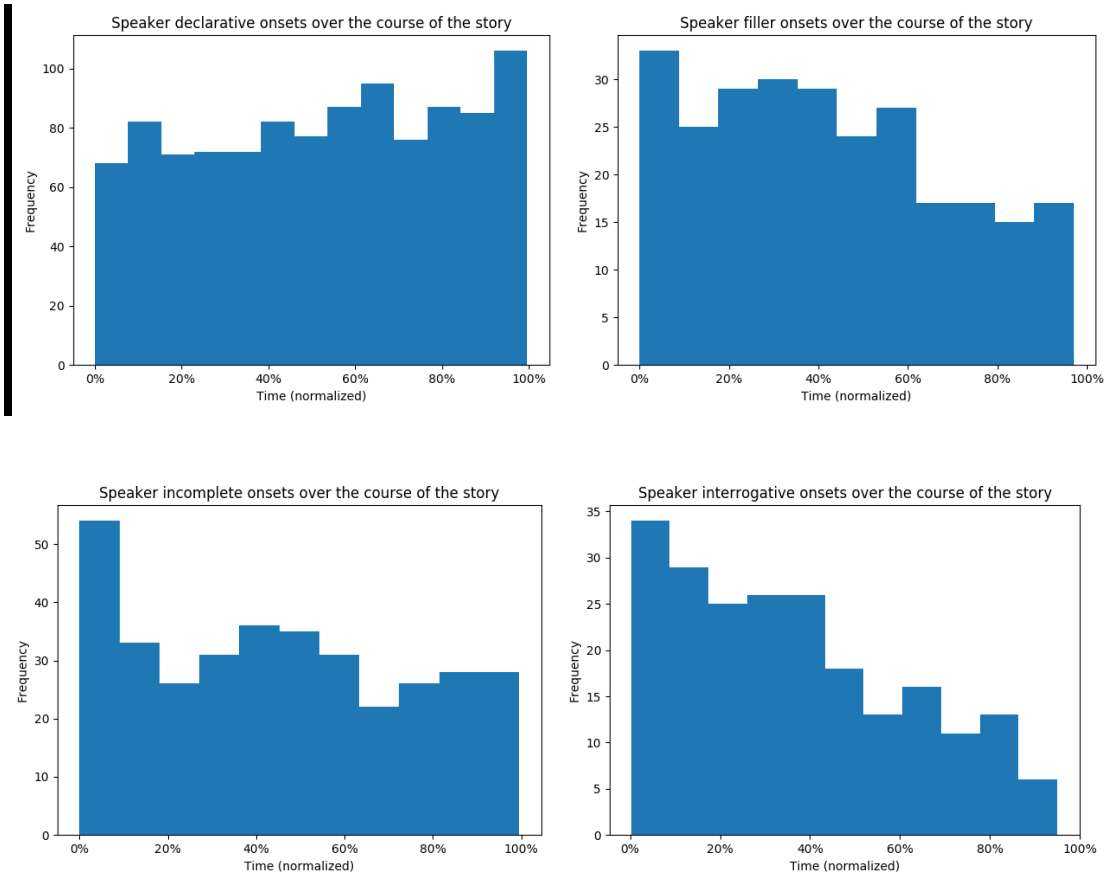
Figure 9. Story histogram – Speaker and Listener back-channel onsets



For back-channels specifically, we see an overall increase in the rate of speaker back-channels as the story progresses, corresponding to the increase in listener declaratives and interrogatives. For listeners we see a slight increase around the 30% mark and a peak after the story has finished, but the rate of listener back-channels is stable for most of the story-telling.

4.1.1 Speech Turns (Speaker)

Figure 10. Story histogram – Speaker declarative, filler, incomplete, and interrogative onsets



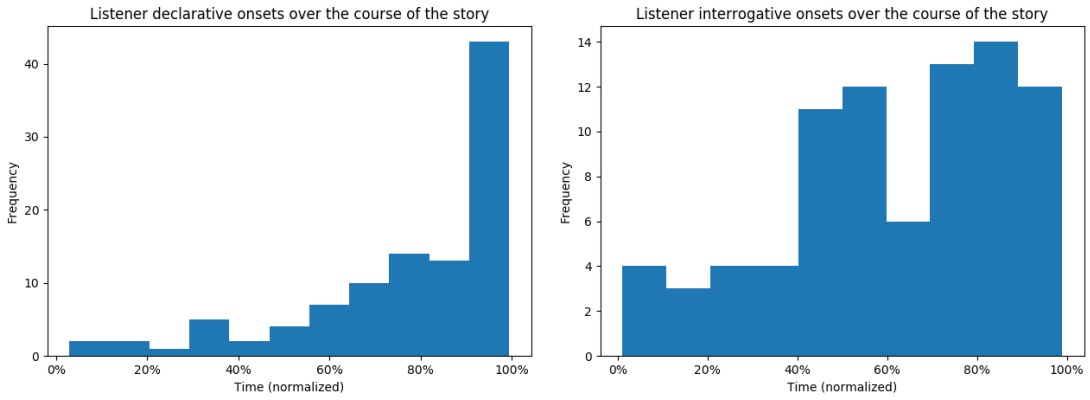
We can see from the distributions of speech turn subtypes that the more or less uniform distribution of speaker speech turns we saw in Figure 8 is in fact quite differentiable across subtypes. Declaratives, which make up the largest subtype of speaker turns, are indeed quite uniform over time, as are incomplete speech segments (except for a peak at the beginning of the story – this is because most of the first speech segments in the stories are not complete clauses). There also seems to be a decline in the usage of fillers at around the 60% mark, approximately when the action is rising higher and the climax tends to start. This may correspond to a more fluid, less pause-filled speech style

designed to engage the listener during the emotional peak, and possibly to increased speech rate. Even after the climax, the rate of fillers does not return to the same height as before.

Interestingly, there is a strong and steady decrease in the frequency of speaker interrogatives over the course of the story. Some of this may have to do with an increase in interrogatives on the part of the listener (see Section 4.1.2), but listener interrogatives don't increase significantly until the climax, while speaker interrogatives begin decreasing almost immediately. The kinds of interrogatives seen in the early parts of the narratives often relate to the setting of the scene, either in directly asking the listener if they are familiar with information that will be relevant for the story (such as whether they know of a particular ski resort or whether they have been to Hawai'i) or, more commonly, in using a question intonation to encourage some sort of response.

4.1.2 Speech Turns (Listener)

Figure 11. Story histogram – Listener declarative and interrogative onsets



There are too few instances of incomplete or filler turns for listeners to plot their frequency distribution. Looking at declaratives, however, we see a small rise during the climax, and then a dramatic spike in the very final moments of the story-telling. The

declaratives spoken during the climax, while the declaratives in the final spike tended to be. The near universal tendency to make such comments seems to be a conventionalized element of this kind of story-telling context, although it would be interesting to know whether this is equally true outside of an experimental context.

4.1.3 Speech Back-channels (Speaker)

Speakers are limited in the number of spoken back-channels they provide, but they are even more limited in their variety of spoken back-channels.

Figure 12. Story histogram – speaker affirmation and laugh onsets

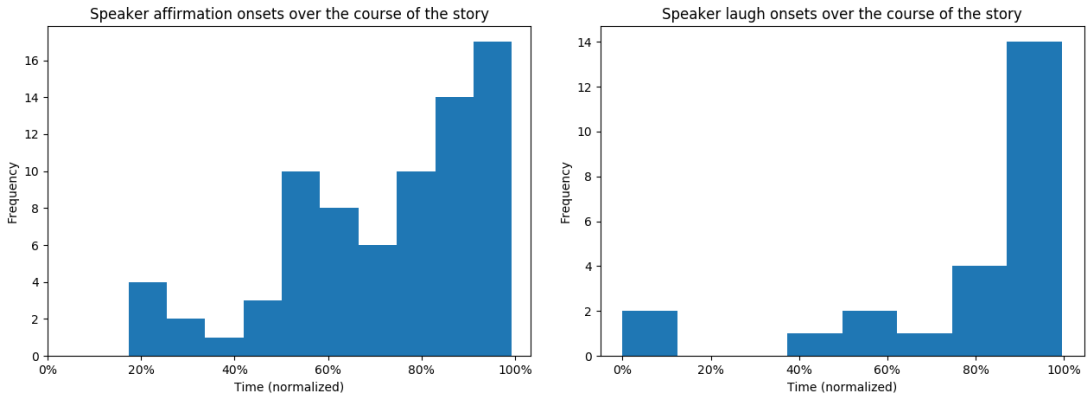


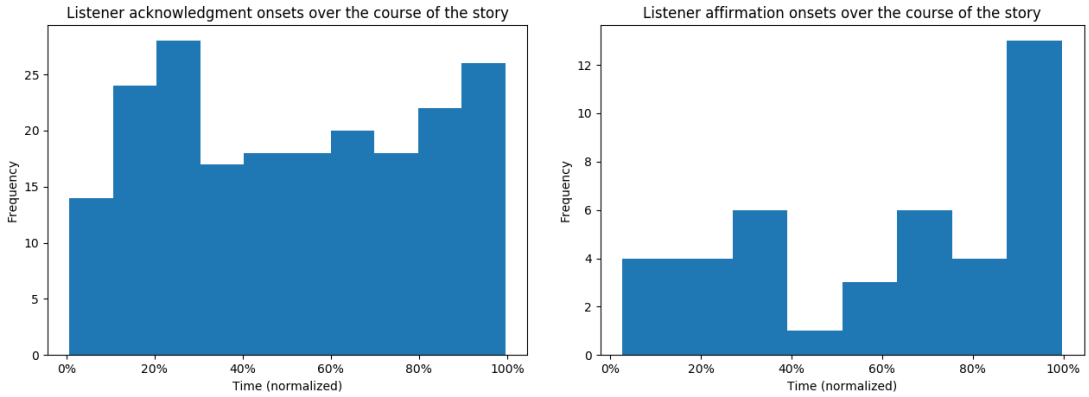
Figure 12 shows the only form of spoken back-channel that speakers reliably use (all other back-channel subtypes had five or fewer tokens). The rate of producing affirmations (mostly *yeah*, *uh huh*, or *right*) tended to increase over the course of the story, the two peaks in the distribution corresponding mainly two the two peaks in the distribution of listener interrogatives (Figure 11), one approximately at the end of the rising action, the other during the falling action and resolution. Figure 12 shows the distribution of speaker laughs, but it should be noted once again that these laughs are laughs which occurred as the only element in a speech utterance, and do not include laughs that were part of a larger speech segment (which occurred throughout the

speaker's story). The laughs in Figure 12, which peak sharply at the end of the story, represent a tendency across most stories for both participants to laugh at the completion of the story, whether because the nature of the stories required a release of tension, or because this was the point at which the style of communication switched from narrative to more turn-based, or some other reason.

4.1.4 Speech Back-channels (Listener)

Listeners exhibited a much wider range of back-channel subtypes than speakers, with a sizable number from seven different subtypes.

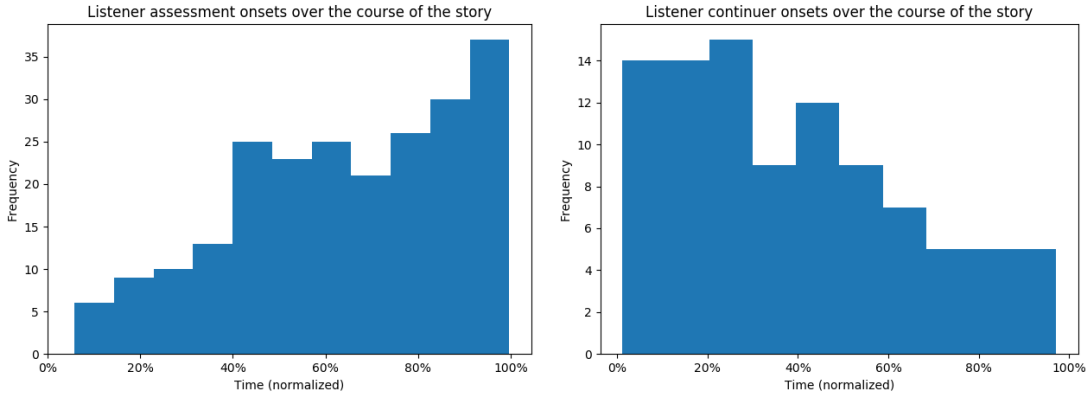
Figure 13. Story histogram – Listener acknowledgment and affirmation onsets



Listener acknowledgements (Figure 13) are used at a fairly consistent rate throughout the story-telling, with a slight increase at the end of the story, and a single peak shortly after the beginning of the story. This back-channel subtype signals acknowledgement of a speaker's speech content, and this first peak in frequency may be an indication that the speaker's setting of the stage is beginning to be acknowledged by a high proportion of listeners. Overall, though, it seems that acknowledging the content of the speaker is a behavior that remains stable throughout the story-telling. Listener affirmations (positive responses to polar questions; Figure 13) peak sharply at the end of the story, where the

nature of the communication becomes more turn-like, and there are a greater proportion of direct questions. Aside from this, however, listener affirmations occur at a relatively stable rate throughout the story, rather than decreasing in frequency like speaker interrogatives. (The majority of speaker interrogatives seem not to be appeals for affirmation, but appeals for evidence of understanding or attention.)

Figure 14. Story histogram – listener assessment and continuer onsets

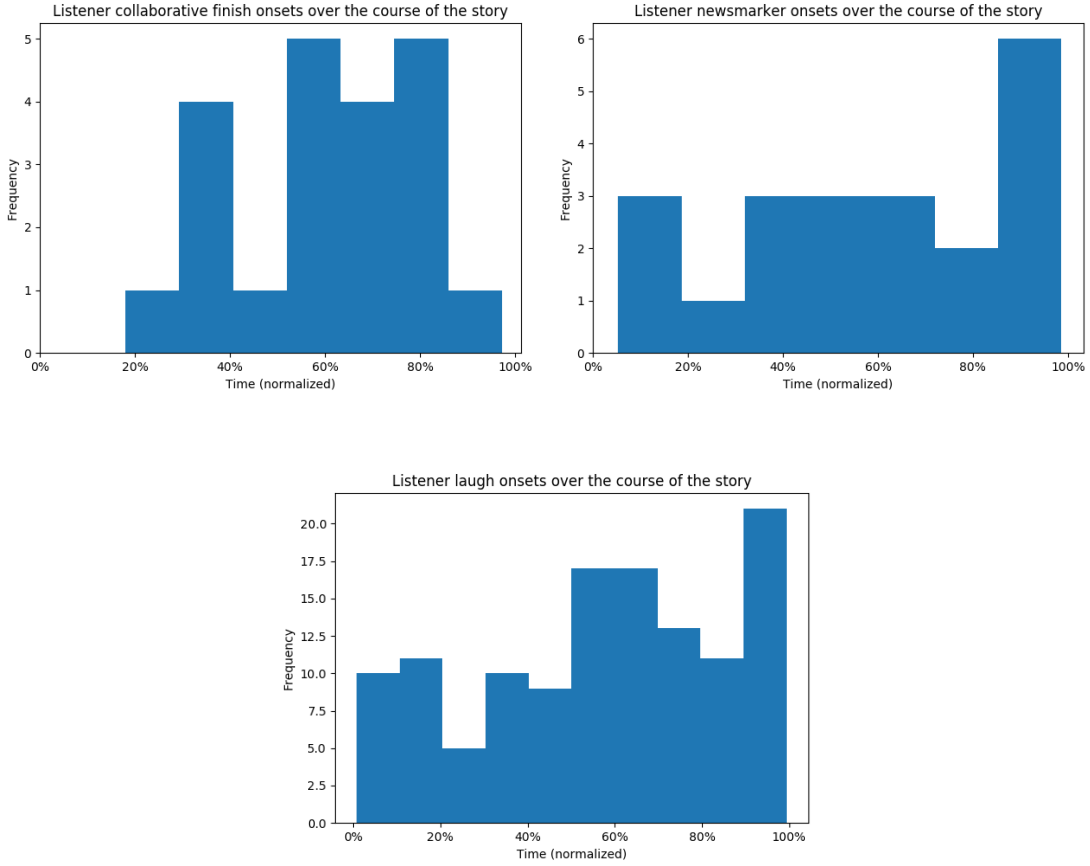


One common type of listener back-channel is the assessment, which expresses some emotional or affective response to speaker content. This can be positive (*so cool!*) or negative (*that sucks!*), or can express emotions like disbelief (*no way!*) or sympathy (*that's scary!*). These back-channels are often in response to especially intense or dramatic parts of the narrative, and we can see that their rate increases steadily over the course of the story, as the action rises and climaxes, and then into the turn-taking portion at the end of the story. Much of this post-story discussion revolves around wrapping up

any loose ends of the story, and a frequent interchange will involve the listener asking a question or making a comment, the speaker responding, and the listener offering an affective response. Assessment back-channels are probably the most effective at expressing interest, and there seemed to be a near universal tendency to let the listener know that the story had been interesting, before moving on to the next task.

Continuers show the opposite pattern. These are back-channels that indicate to the speaker that they should continue speaking, without clearly acknowledging the content of the speech. These back-channels occur frequently at the beginning of the story, then drop off in frequency as the story progresses. This may be for multiple reasons. It may be that continuers drop off in frequency because the beginnings of the stories tend to hold less dramatic information, and as the drama rises, assessments take the place of continuers. Or it may be that continuers are a more appropriate response to speaker interrogatives (or some types of speaker interrogatives), and so they decline as the rate of interrogatives declines.

Figure 15. Story histogram – Listener collaborative finish, newsmarker, and laugh onsets



Collaborative finishes and news-markers are fewer in number than other listener back-channels. There is not much to be drawn from this data except that collaborative finishes tend to occur during the middle of the stories rather than the ends, and that news-markers are relatively consistent throughout the story, with a rise at the story's end. Laughs are also fairly consistent throughout, except for an apparent peak around 50-70% of the way through the story, occurring around the time that the climax is being revealed, and a sharp peak at the end of the story, as participants are likely to be responding with a laugh to the story as a whole or to the speaker's comments on the story. Besides being a response to humor, laughter functions to release tension and build rapport (Strean 2009, Jefferson, Sacks, & Schegloff 1987). The fact that these listeners are listening to life or death

stories, which are full of tension, may be influencing the rate and function of the laughter seen here.

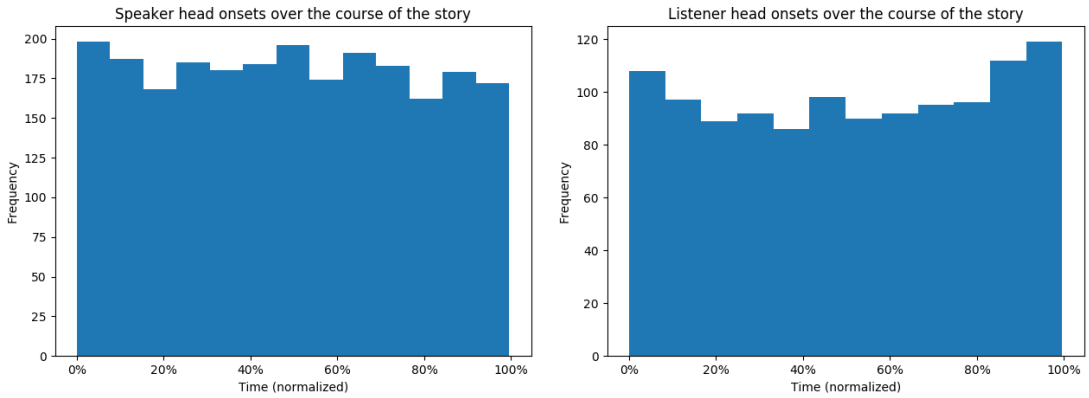
An important finding of this analysis is that while the rate of different subtypes of listener back-channels change, often dramatically, over the course of the story-telling, the rates of speaker speech turns and overall listener back-channels do not (Figure 8). For the majority of the story-telling, listeners are consistent in their rate of back-channels. This suggests two possibilities, offered here as extremes. It may be that back-channels occur entirely independent of the interlocutor: regardless of speaker behavior, the social convention in this story-telling context may be to give a constant stream of evidence of attention, and so to back-channel at a consistent rate. Or it may be that back-channels are entirely dependent on the interlocutor: regardless of how long it has been since the listener has performed a back-channel, they will only do so in response to certain configuration of speaker behavior. Of course, it is likely that listener back-channels have varying degrees of dependence on and independence from speaker behavior. We will look at this in more detail in Chapter 7, Section 3.

4.2 Head Gestures

Head gesture is also an area in which we have a variety of subtypes to analyze. The number of subcategorizations in this modality is large enough, in fact, that there are too few tokens of some subtypes to identify meaningful patterns. For this reason, we will focus on the larger categories of shakes, tilts, and other categories.

4.9a-b: Story histogram – Speaker and Listener head onsets

Figure 16. Story histogram – Speaker and Listener head onsets

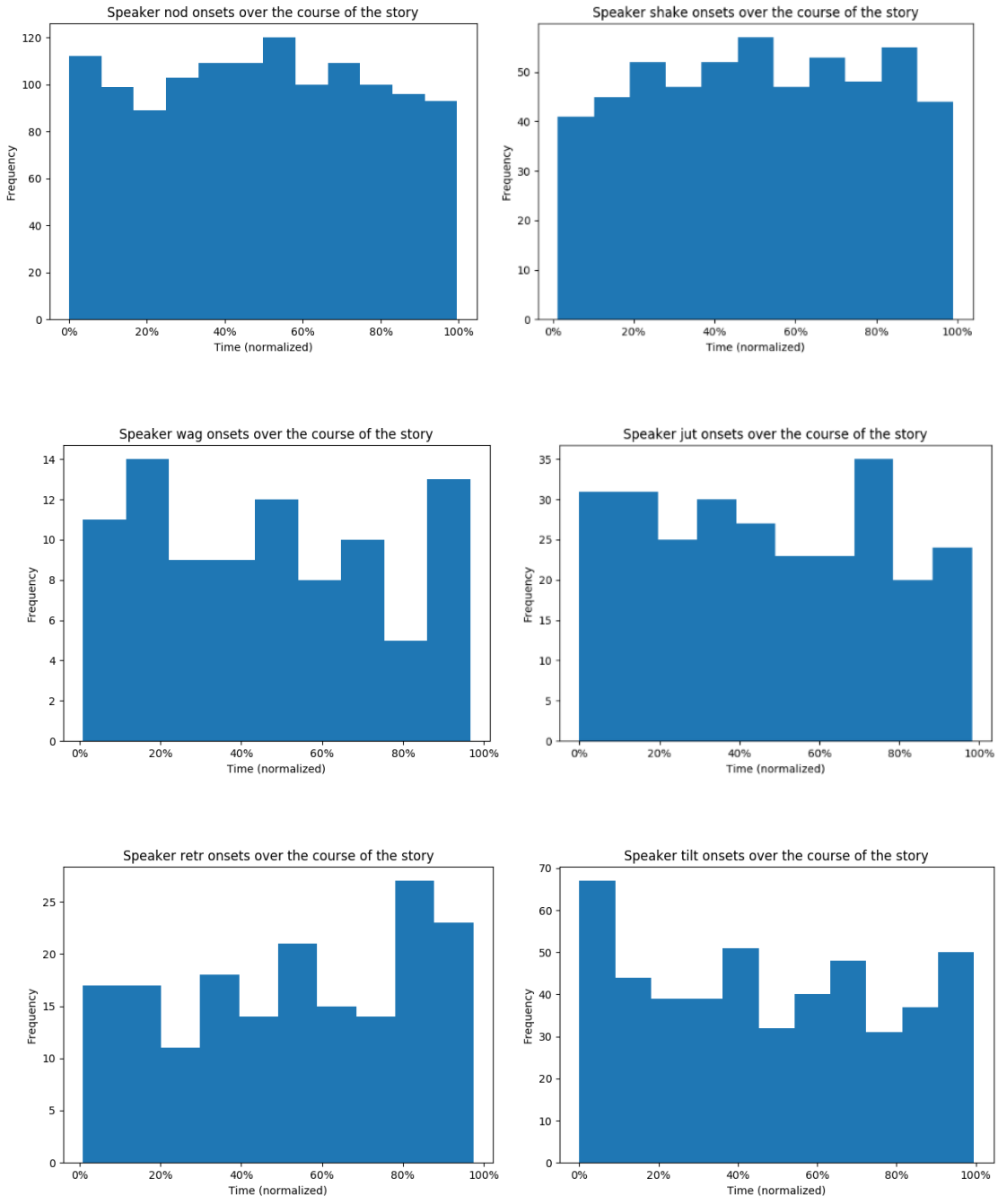


Looking at all head behaviors together (Figures 16), we see very little differentiation across roles. Like with spoken turns and spoken back-channels, both listeners and speakers are engaging in head gestures consistently throughout the story-telling. There is a slight increase in listener head gestures at the end of the story, around the time that the communicative style shifts to be more interactive.

4.2.1 Speaker Head Gestures (Axial Subtypes)

Speakers use head gestures around twice as often as listeners, and their gestures are more evenly distributed across the different subtypes, so there are sufficient tokens of each subtype to display their distribution here.

Figure 17. Story histogram – Speaker nod, shake, wag, jut, retraction, and tilt onsets

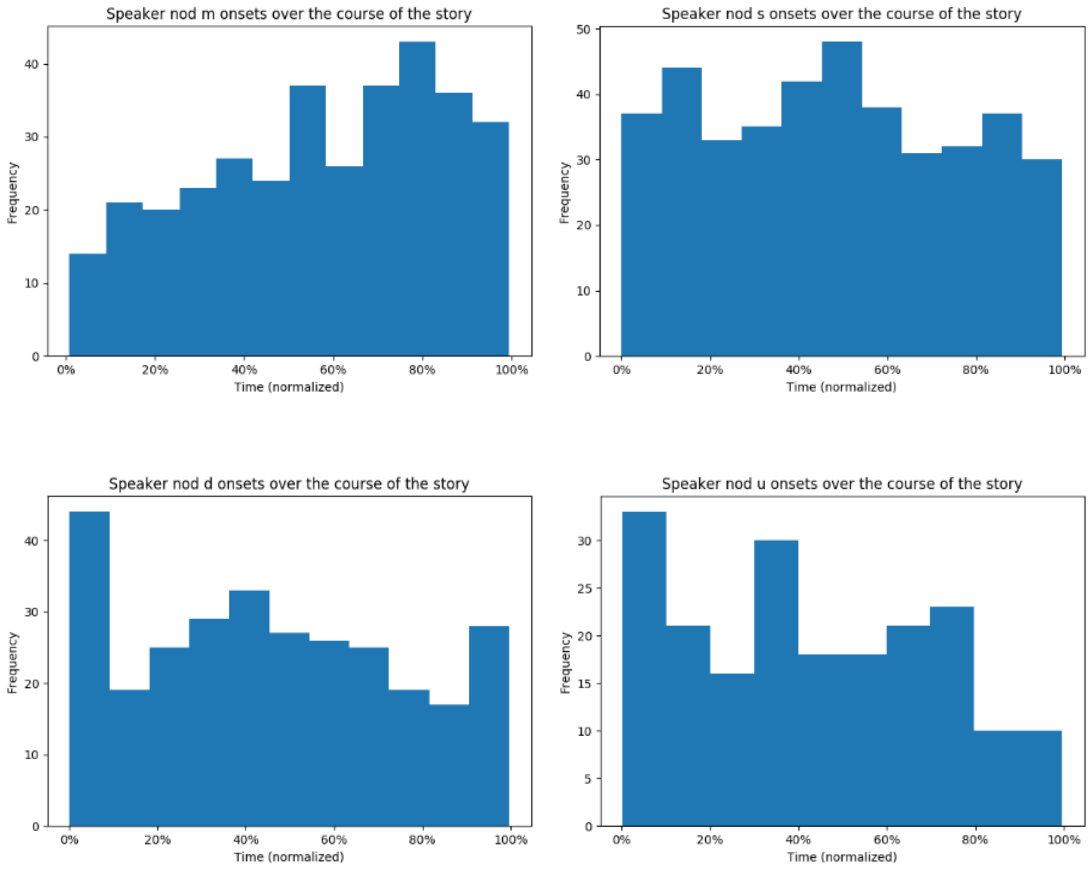


However, in Figures 17, we little variation in the subtypes across the course of the story-telling worth discussing. Nods and shakes remain consistently in use throughout. Wags, juts, and tilts show some suggestion of a decreasing over the course of the story, and

retractions of increasing, but the number of tokens is too small and the slopes too uncertain to draw any conclusions. There are a couple of notable patterns at the extremes of the distributions, however. Retractions peak at the end of the story, and this may be tied to the turn-taking that occurs then, but for some participants this gesture seems to be used at the end of the story (often accompanying a *yeah*, or *so that was crazy*) as a way of remarking on the intensity of the story they have just told. Tilts have a peak at the beginning of the story, and in these cases seem to have the function of setting up a shift into a story-telling perspective or stance.

4.2.2 Speaker Head Gestures (Axial + Cyclic Subtypes)

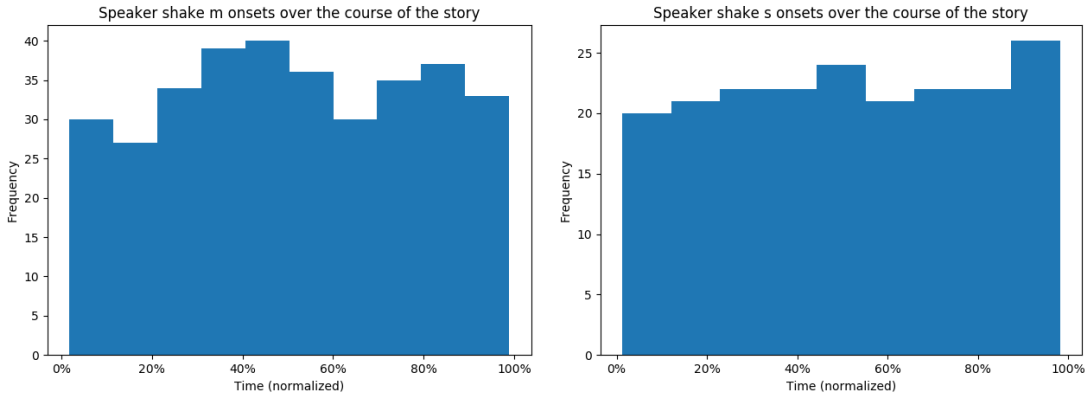
Figure 18. Story histogram – Speaker multiple nod, single nod, nod down, and nod up onsets



Speaker nods (Figures 18) show some differentiation when broken down further into different cyclic patterns. Multiple nods tend to increase over the course of the story, peaking during the climax of the story. This aligns with findings in the literature (Chapter 1, Section 5) suggesting that multiple nods have the function of emphasizing content, the climax being the section of the story where emphasis is most appropriate. Note that single nods do not follow this pattern, suggesting that they do not always have the same function as multiple nods, although they have a formal similarity: the nadirs of single nods also tend to co-occur with the stress of a word in the speech stream.

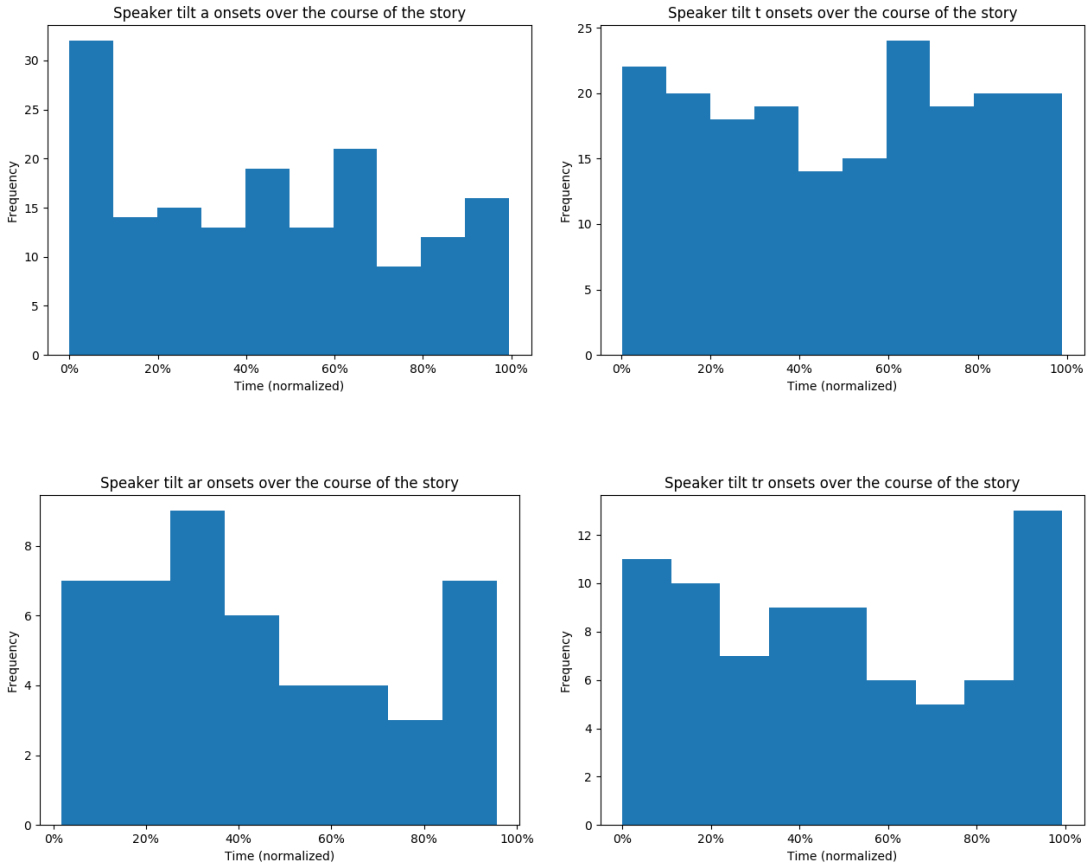
Nods-down may have some increase in frequency during the rising action, but they certainly have a peak at the very beginning of the story. These story-initial nods-down seem to function similarly to story-initial tilts, as a way of shifting into a story-telling mode (these frequently co-occur with the *so* introduction). We see a similar peak in nods-up at the beginning of the story, which may have a similar function. These head gestures show a decrease over the course of the story, which may be related to the decrease in speaker interrogative speech segments, which these tend to co-occur with.

Figure 19. Story histogram – Speaker multiple and single shake onsets



By contrast, as we can see in Figures 19, the cyclicity of speaker head shakes does not seem to be a differentiating factor like it is for head nods. It may be that single and multiple speaker head shakes are not functionally different behaviors, although this only negative evidence, so this is only a conjecture.

Figure 20. Story histogram – Speaker tilt away, tilt toward, tilt away + return, and tilt toward + return onsets



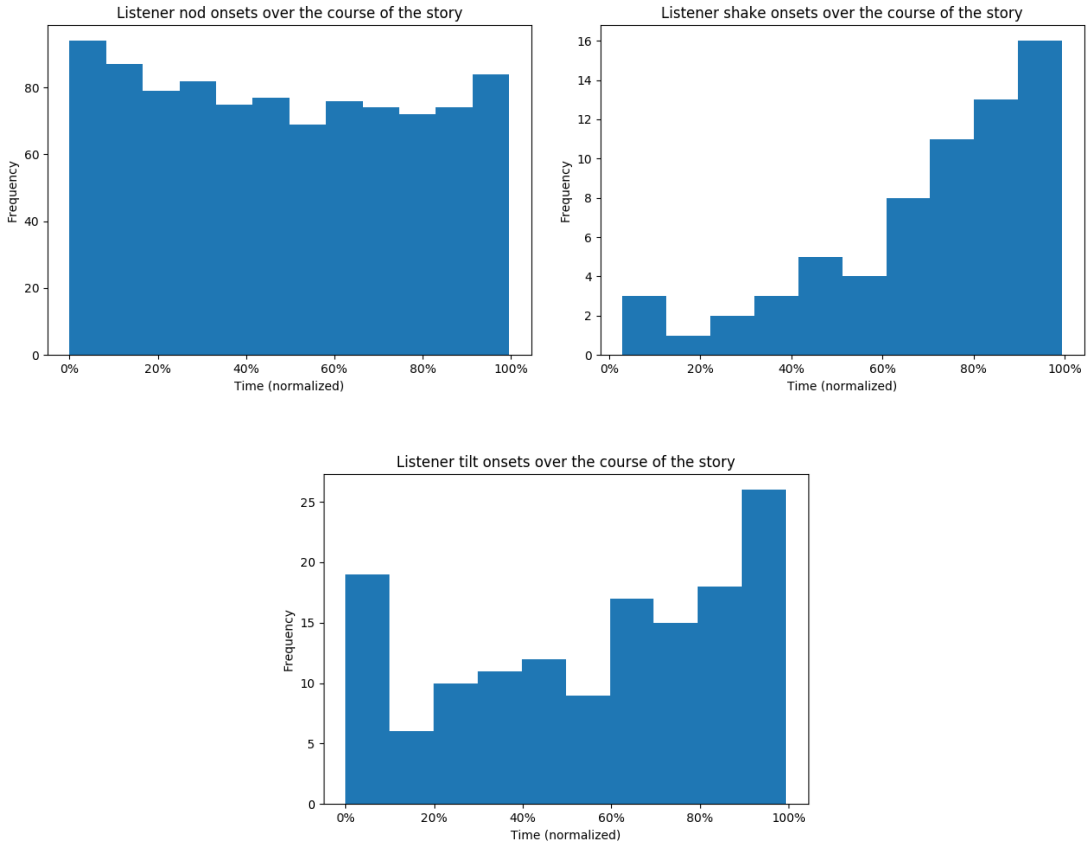
Looking at the cyclic subtypes of the tilts (Figure 20), we can see that the story-initial peak in head tilts comes primarily from tilts-away, which are otherwise relatively consistent throughout the story[20]. The number of tokens is not large enough to make confident claims about apparent slopes or peaks for other behaviors, aside from the peaks in cyclic tilts (Figure 20) at the close of the story.

[20] Interestingly, while all dyad types often start out with a tilt-away, this head gesture rapidly decreases over the course of the story for everyone except women telling stories to men.

4.2.3 Listener Head Gestures (Axial Subtypes)

There are too few tokens of listener juts, retractions, and wags to show the frequency distributions, so we will only examine listener nods, shakes, and tilts.

Figure 21. Story histogram – Listener nod, shake, and tilt onsets



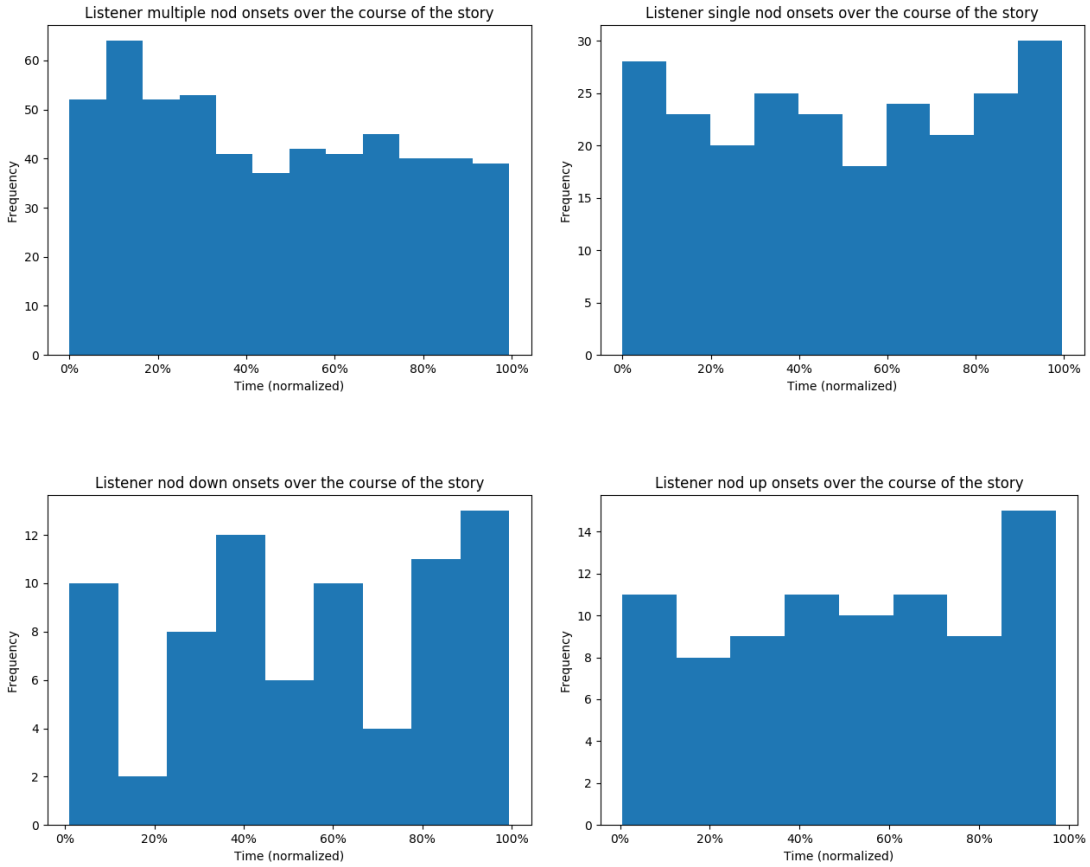
Unlike with speakers, we see in Figure 21 that listener head gestures are very clearly differentiated across subtypes. On the one hand, listener nods are used consistently across the course of the story. Listener shakes, on the other hand, occur very infrequently during the first half of the story, and increase rapidly in frequency during the climax and resolution. These listener head shakes tend to occur in response to events in the story that are negative in some way, such as the revelation that the speaker got hypothermia, or nearly flipped their car. In fact, they have similar function to the assessment back-channel

in the speech modality, although the emotional responses they tend to express are sympathy and disbelief. Listener tilts also increase over the course of the story. It is not clear what the function of these tilts is, although it may be that they have the general function of indicating a shift of perspective, which for the listener can indicate the processing of the speaker's speech. Listener, like speakers, also show the same peak of tilts at the beginning of the story, a shift into story-listening mode, possibly mirroring speakers, or possibly also signaling that they are prepared to attend to the story.

4.2.4 Listener Head Gestures (Axial + Cyclic Subtypes)

Only listener nods had sufficient tokens to look for different patterns in cyclicity. However, as we can see in Figure 22 below, it does not appear that different subtypes of listener nods are meaningfully different across the course of the story, except for a tendency for listeners to produce more multiple nods towards the beginning of the story.

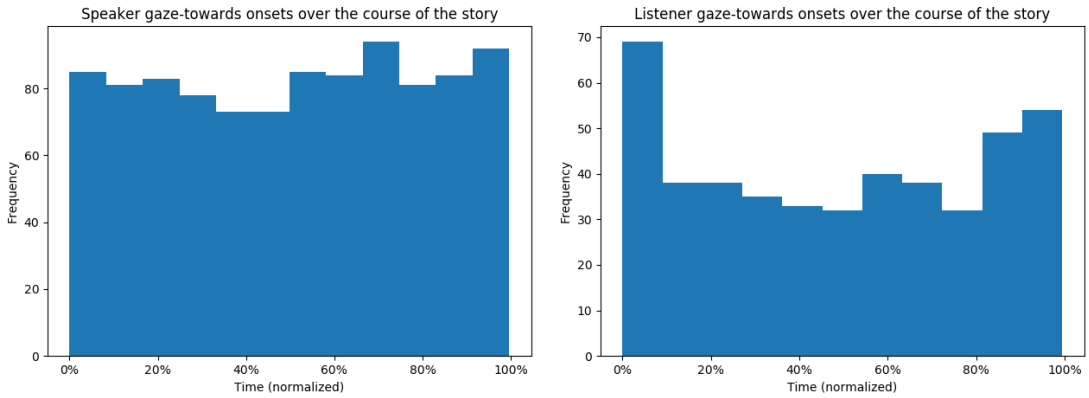
Figure 22. Story histogram – Listener multiple nod, single nod, nod down, and nod up onsets



4.3 Speaker and Listener Gaze

Gaze-shift is fairly uniform throughout the stories, for both listeners and speakers (Figure 23). Listeners shift their gaze towards speakers at the beginning of the story, resulting in a frequency peak, and their rate of gaze-shift peaks again as the story winds down, but aside from that, they shift gaze at a relatively uniform rate. Speakers' rate of gaze shift doesn't change at all across the course of the story.

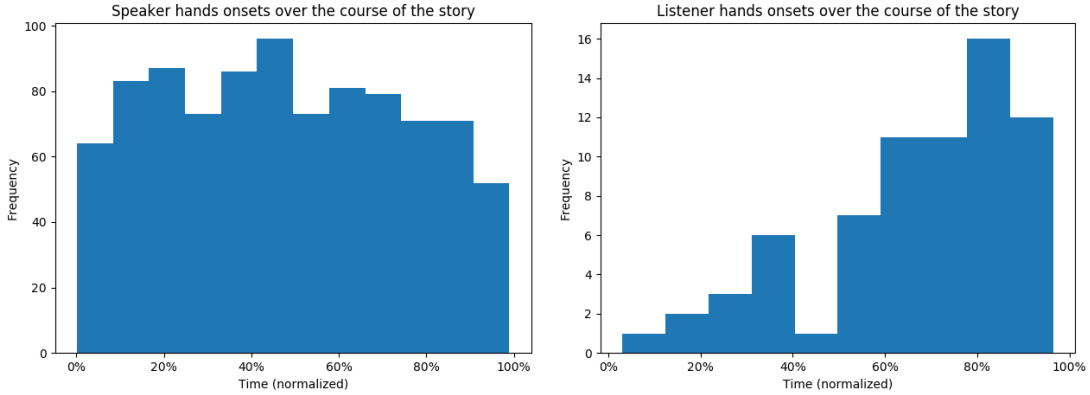
Figure 23. Story histogram – Speaker and Listener gaze-towards onsets



4.4 Speaker and Listener Manual Gesture

For listeners, the rate of manual gesture increases steadily over the course of the story (Figure 24), which follows the increased rate of speech turns and back-channels, seen above in Figure 9, since listener manual gesture is co-speech gesture. For speakers, the rate of manual gesture remains relatively uniform throughout. One might have expected that it would increase around the climax, but there doesn't seem to be any such correlation. This is one of many points where we see no quantificational evidence for a common communicative function of gesture, but rather a fairly constant use of gesture with speech production independent of speaker goals/purposes.

Figure 24. Story histogram – Speaker and Listener manual gesture onsetes



5. Summary and Possible Hypotheses

5.1 Summary

Section 2 of this chapter looked at the lag times between modalities for listeners and speakers. We saw that the lag times between modalities were mostly inversely related to the durations of those same modalities – behaviors with long durations tended to have shorter lag times between. This was not the case for gaze shift, however, which differs from the other three modalities in only requiring effort to shift, not to maintain. Both listeners and speakers had relatively short average durations of gaze-away, but speakers looked away much more frequently. Lag times between head gestures were also fairly similar across roles, another way in which head gesture appear relatively similar across roles.

Section 3 looked at bigrams of within-role, within-modality behavior segments of head and speech subtypes. The strongest dependencies in each modality were bigram pairs of the same subtype (e.g. multiple nod + multiple nod or assessment + assessment). Wags and juts (and retractions, for speakers) were not strongly dependent on other head

behaviors. For listeners, there was a strong pattern of bigrams consisting of a nod followed by a half-cycle, repositioning head gesture.

Section 4 looked at frequency distributions (*story histograms*) of individual behaviors, collapsed across all normalized timelines of all stories in the corpora. In the speech modality, it was found that, while speakers' rate of speech overall is relatively uniform over time, different subtypes varied substantially. Speaker interrogatives decreased over the course of the story, while speaker back-channels increased, and speaker fillers diminished during the climax. For listeners, rates of back-channels increased slowly, and rates of speech turns increased dramatically at the end of the story. Listeners' back-channels also showed interesting variation, with assessment back-channels becoming more frequent over time and continuers becoming less frequent.

In the head modality, both speaker and listener head rates overall were uniform. Speaker head subtypes didn't show strong variation over time, although rates of tilts were very high at the beginning of stories. Listener head subtypes showed more differentiation, with nods remaining uniform over time, but shakes increasing, in a similar fashion to listener assessments.

5.2 Hypotheses

There are a number of hypotheses one could formulate from the data in this chapter. A small sample is laid out below.

1. There may be many reasons why head gesture exhibits the greatest similarity across roles, of all modalities examined here. 1) Listener frequency approaches speaker frequency because head gesture doesn't impede the speaker's message like listener speech would; 2) the 'window of opportunity' for head gesture

(interlocutor gaze) is longer, overall, than the 'window of opportunity' for speech (interlocutor speech pause); or 3) listener head gesture responds to or is somehow dependent on speaker head gesture.

- The first explanation would be difficult to test directly (we cannot manipulate how much one person's head gesture impedes another's).
- The hypothesis for the second explanation is that people produce more head gestures when they are easily detectable (i.e. easily seen). This could be tested by comparing individuals who had longer and shorter total gaze-time towards their interlocutors, and looking for a positive correlation between longer gaze-times and greater rate of interlocutor head gesture. The same could be done for speech, in fact, comparing individuals with longer or shorter proportions of speech-time during the story, and correlating these with proportions of the interlocutor's spoken back-channels.
- The hypothesis for the third explanation is that one interlocutor's head gesture is dependent on the other's. This could be tested by looking at correlations of proportions of each interlocutor's head gesture, with a positive correlation being evidence for this hypothesis.

2. The average lag time between listener speech segments was 50% longer than the average lag time between listener head segments. Does the rate one of these behaviors contribute more than the other to an onlooker's perception of listener comprehension, rapport, or responsiveness, and would changes in the rate of each modality shift these perceptions equally?

- One set of hypotheses is that a participant's rate of speech (or head gesture) influences onlookers' judgment of that participant's level of comprehension, rapport, or responsiveness. This could be tested by selecting two sets of video clips from storytelling elicitations: one set with a higher rate of speech or head gesture, and one with a lower rate. Experimental participants would watch the video clips and rate the speakers and listeners in these clips on these three target judgments. A prediction might be that a greater rate of speech and gesture would correlate with higher ratings on all these judgments for both speakers and listeners.

3. A nod followed by a half-cycle repositioning head gesture is a highly frequent bigram for listeners. Do these bigram constructions differ from a solitary listener nod in terms of the content of the speaker speech it co-occurs with?
    - We might hypothesize that the nod + half-cycle bigrams will be more likely than the unigram nods to occur near specific linguistic features, such as clause breaks, discourse connectors, or filled pauses, and test this by finding the odds ratio (see Chapter 2.6) of the overlap of these n-grams with the target speech segments to see whether these overlaps are more or less likely than expected, and comparing them for the bigrams and unigrams. We could also look at discourse pragmatic features, such as the introduction of new events or referents, or the use of perspective-shifting linguistic forms (e.g. *however, on the other hand,* etc.). Greater than likely overlap between the bigrams and such linguistic segments is one way to provide evidence for a functional interpretation of such a head gesture construction.

4. The most dependent listener back-channel bigrams are repeated back-channels (such as assessment + assessment and acknowledgment + acknowledgment). In these bigrams, is each unit responding to the same topic, or to different topics? It's not entirely clear when and why listeners produce back-channels. Certainly they often respond to the content immediately prior, but are they also influenced by the entirety of the content since their last back-channel, or by the nature of their own last back-channel?
    - The fact that the three most dependent listener back-channel bigrams are repetitions suggests two possibilities: 1) the content of the speaker's speech immediately prior to the back-channels is similar in nature (and so elicits a similar back-channel), or 2) the second back-channel unit in these bigrams is motivated by the same considerations as the first, rather than the immediately preceding speaker's speech.
    - Both of these hypotheses could be tested by examining repeated back-channel bigrams in a corpus and looking at the preceding speaker speech for each bigram unit. If, for most bigrams, the preceding speaker content is similar before each bigram unit, this would suggest that these repetition bigrams are driven by the immediately prior speaker behavior. If, for most bigrams, the first speaker content is appropriate (by some standard) and the second speaker content is not, this would suggest that both back-channels are driven by the first speaker content.
5. Speakers use more interrogatives towards the beginning of their stories. Are these being used in functionally different ways that change over the course of the story?

- One hypothesis might be that, at the beginning of the story, listeners use interrogatives to establish rapport, or demonstrate that they care about or are invested in the storytelling. Another might be that, at the beginning of the story, there is more that is unknown, and so listeners ask more questions to fill in these information gaps.
- Either of these factors could be experimentally manipulated, and the proportion of early-story interrogatives could be compared across conditions. For the first hypothesis, the impetus to exhibit investment or establish rapport could be manipulated by comparing strangers' interactions (as in this dataset) vs. friends (whose rapport would already be established), or by having the participants tell multiple alternating stories (so investment would demonstrated more in the earlier stories, although a decline in the need to demonstrate investment might be conflated with a fatigue effect). If early-story interrogatives were equally frequent for both conditions, this would suggest that their function is information-gathering.
- For the second hypothesis, the impetus to fill in information gaps could be manipulated by having one group of participants listen to stories from a stranger that they had already heard or read before (but they are not meant to let the speaker know this), and the other group listen to stories that were novel to them. If early-story interrogatives were equally frequent in both groups, this would suggest that their function is rapport-building or investment-exhibiting.

6. The narrative structure of these stories is quite specific. Would other kinds of stories show the same patterns (such as increased rates of assessments and listener shakes during rising action and climax)?
    - The underlying hypothesis here is that larger discourse structures influence the use of multimodal behaviors.
    - More specific hypotheses could be made about the stability and instability of different interactions between the structure of a communicative context and the use of multimodal behaviors. For instance, one might hypothesize that assessment back-channels will grow more frequent across all communicative contexts (because the urge to show interest increases as the dialogue drags on) or that this is specific to certain kinds of storytelling (because they follow the rising action and climax of the narration). This could be tested by comparing the frequency distributions of assessments in multiple communicative contexts, such as storytelling, instructions-giving, and persuasive arguments (although it may be that non-storytelling contexts also have rising action and climaxes).
7. How dependent or independent of speaker behaviors are back-channels? Will manipulating the visibility of speaker cues influence the rate of back-channels?
    - We can hypothesize that listeners back-channel produce back-channels in response to certain speaker cues, such as gaze direction or pitch contours in speech (in addition to the semantic content of the speaker's speech).
    - To test these, we could obscure these cues and compare the timing of listener back-channels in an obscured-cue group and a non-obscured-cue group. This

would be relatively easy for gaze – simply put a screen between participants or have them speak with each other remotely. We might predict that head gesture back-channels would be less frequent, but we might predict that the overall rate of back-channels (including spoken back-channels) would not differ. (We could also manipulate whether or not the speaker could see the listener, such as by blindfolding only the listener, to see whether speaker gaze is influential.)

- For speech, we could use software to resynthesize the speaker's pitch to a constant frequency, eliminating an important prosodic cue to speech offsets (although not syntactic or semantic cues). Here, we might predict a reduction in spoken back-channels, because the window would be less reliable, but not a reduction in back-channels overall.

CHAPTER V: WITHIN-ROLE / ACROSS-MODALITY

1. Introduction

This chapter looks at timing relationships that exist across different modalities, but within the same individual. The structure of this chapter will be similar to the structures of Chapters 6 and 7, with one section providing an overview of the modalities of interest, and the rest of the sections being pairwise comparisons of the modalities, looking at different kinds of likelihood measures, n-grams, and window histograms[21]. In this chapter, section 2 is an overview of the modalities, looking broadly at timing relations between modalities for speakers and listeners. Section 3 looks in detail at the timing relations between subtypes of speech and head gesture. Section 4 looks at different head types and their timing relations with gaze towards the interlocutor (and, by extension, gaze-away). Section 5 looks at the timing relations between speech types and gaze-towards. Finally, section 6 will summarize the findings and suggest some hypotheses, arising from the results, which could be tested with qualitative analysis.

2. Overview of the Four Modalities

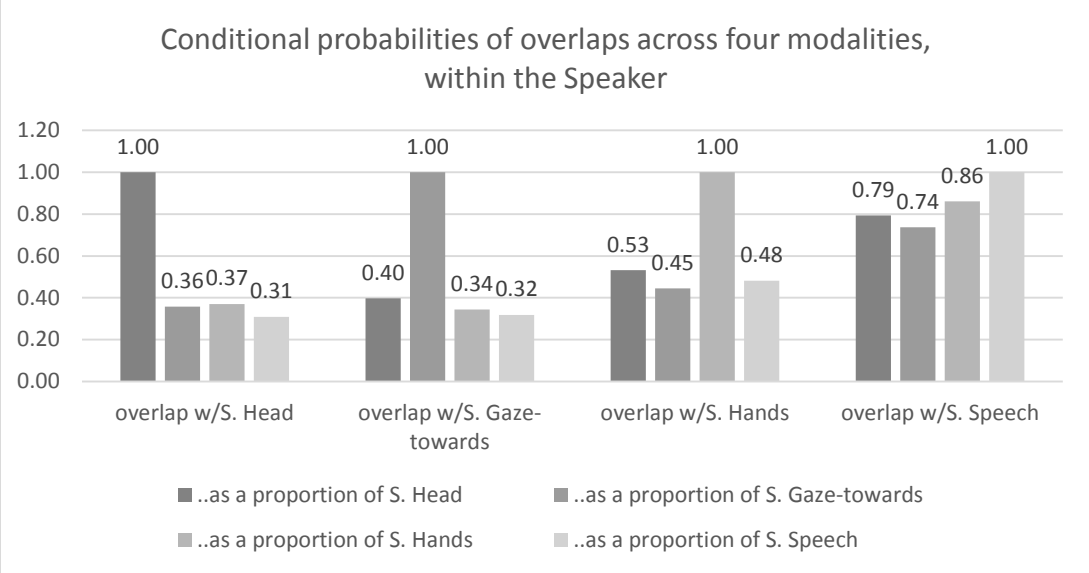
2.1 Likelihood Measures

To begin, we will look at the overlap of each of the four modalities, without breaking heads and speech into subcategories, to get a broad picture of the nature of their dependencies. Figure 25 below shows the conditional probabilities of each speaker modality with each other speaker modality. To give some examples, the left-most bar shows the overlap between speaker head and speaker head as a proportion of speaker head (100%, of course); the next bar shows the overlap between speaker head and

[21] Because very few interesting interactions were found between hands and gaze, hands and speech, and hands and head, these pairwise analyses are not included in this chapter.

speaker gaze-towards, as a proportion of speaker gaze-towards (36%) – thirty-six percent of the time that speakers are looking at listeners, they are also producing head gestures.

Figure 25. Conditional probabilities of overlaps across four modalities, within the Speaker

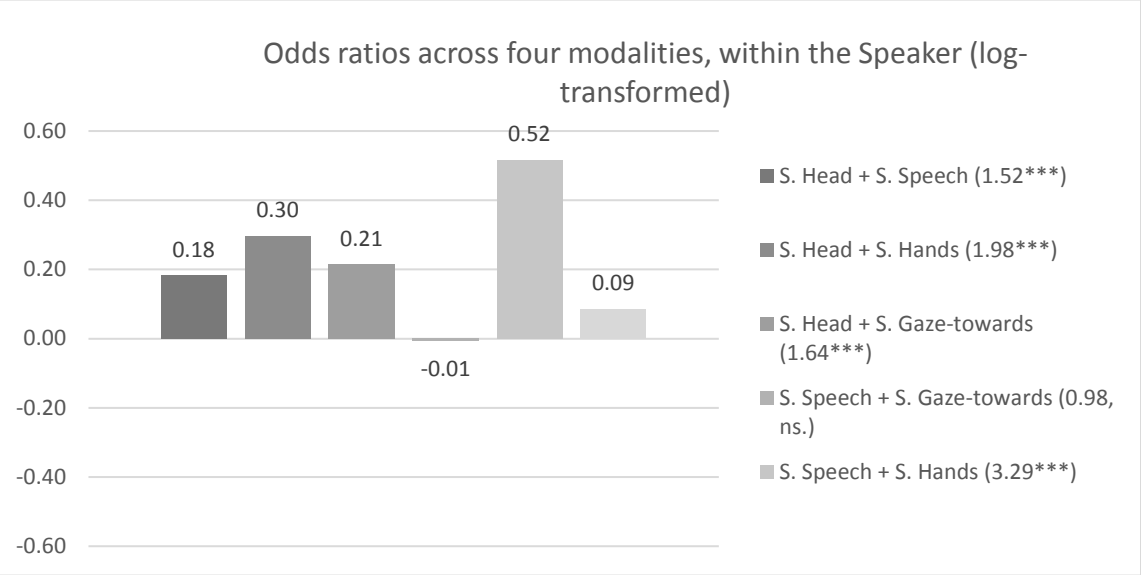


These overlap values measure the number of frames in which both modalities are being produced as a proportion of the total number of frames for each individual modality. In the first set of bars, we see that speaker head gesture is being produced during about a third of each other modality. Speaker gaze-towards also overlaps with around a third of other modalities, up to 40% of speaker head gesture. Speaker hands show somewhat more overlap, co-occurring with around 50% of each other modality. Speaker speech shows the most overlap, co-occurring with around 80% of other modalities.

Another way to look at the dependencies across modalities is to examine the ratio of observed overlap to the overlap that would be expected from the random distribution of the two behaviors, given their relative frequencies. The odds-ratios of these overlaps can be seen in Figure 26, and will be referenced frequently throughout this chapter and

Chapters 6 and 7 (see Chapter 2, Section 6 for a discussion). For the purposes of visualizing, the odds-ratios below have been log-transformed, but the original odd-ratios are included in the legend[22].

Figure 26. Odds ratios across four modalities, within the Speaker (log-transformed)

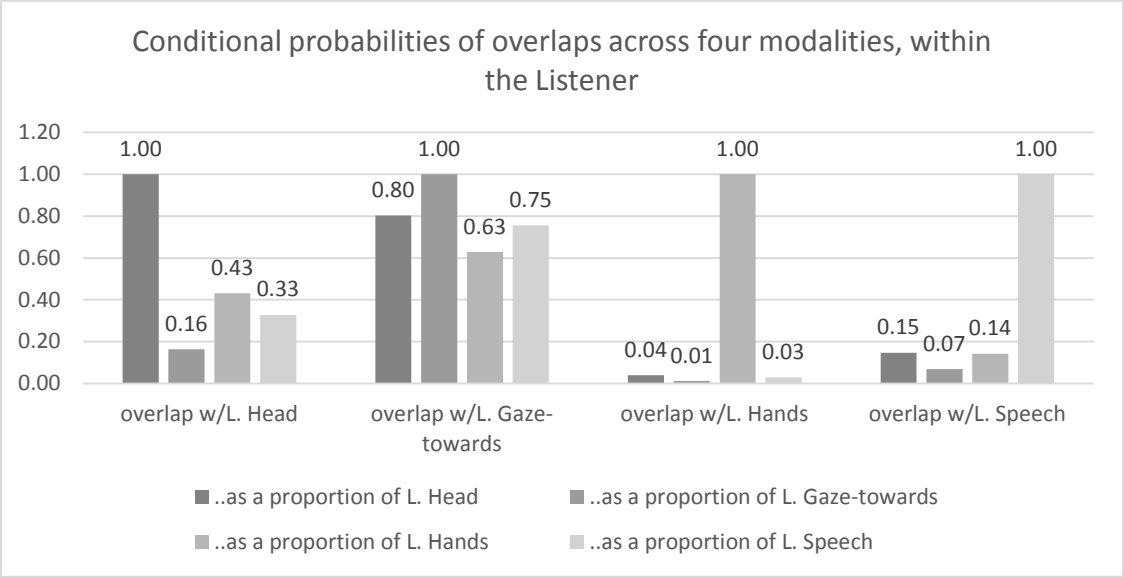


We see here that most overlaps between modalities are more likely than we would expect from chance. The greatest positive differences are in the overlaps between speech and manual gesture (3.29, more than three times more likely than expected) and between head and manual gesture (1.98, almost twice as likely), unsurprising given that manual and head gesture are typically co-speech. More interesting is that head gesture overlaps more than expected with gaze-towards the listener (1.52), while speech is at chance (0.98), and manual gesture is only slightly more likely (1.22). One possibility for this is that speaker head gesture may be sometimes intended to elicit a response, and so the speaker looks to check that the signal has been delivered.

[22] For all odds ratios, the conventions for expressing whether the observed proportion is significantly different than the expected proportion are as follow: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

In Figure 27 below, we see the conditional probabilities between the same four modalities in listeners, and these look very different.

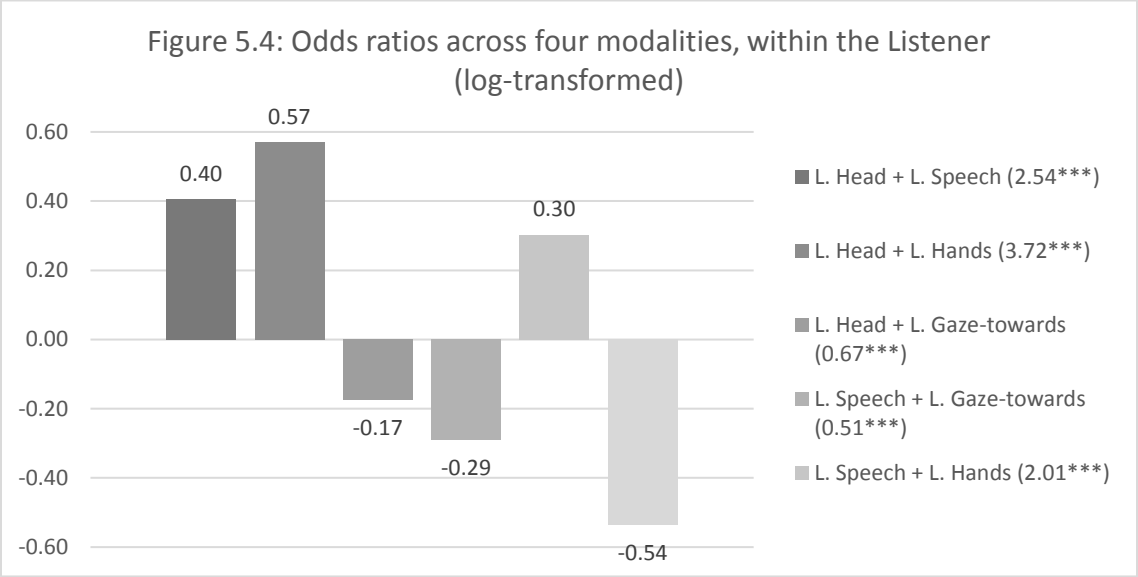
Figure 27. Conditional probabilities of overlaps across four modalities, within the Listener



Listener heads overlap with hands and speech to a similar degree as speakers, but their overlap makes up a much smaller proportion of listener gaze-towards (16%), about half the proportion see in speakers. Listener gaze, on the other hand, overlaps with a large proportion of other modalities (63-80%) – more for head gesture than manual gesture. Listener hands are both infrequent and short, and don't account for much overlap with other behaviors. Listener speech also doesn't overlap with much of the other behaviors, at least not as much as head gesture.

Figure 28 shows fewer overlaps that are more likely than expected than we saw for speakers in Figure 26.

Figure 28. Odds ratios across four modalities, within the Listener (log-transformed)



Speech, manual gesture, and head gesture all display more than expected overlap, each of these often co-occurring in listener spoken responses. But listener gaze-towards the speaker overlaps much less than expected with all three other behaviors, suggesting that listeners are often looking away to produce these other three behaviors together.

2.2 N-grams

We start by looking at bigrams of onsets of all four speaker behaviors, in Table 10.

Table 10. Speaker onset bigrams by modality (1-second window)

Bigram (1 + 2)	Frequency	Symmetric CP	CP: 2|1	CP: 1|2
Gaze-towards + Head	367	0.062	0.163	0.379
Speech + Head	473	0.054	0.209	0.256
Head + Gaze-towards	341	0.053	0.352	0.151
Head + Gaze-away	325	0.050	0.345	0.144
Gaze-away + Speech	287	0.047	0.155	0.304
Hands + Head	282	0.041	0.125	0.329
Head + Speech	414	0.041	0.224	0.183
Head + Hands	248	0.032	0.290	0.110
Speech + Hands	215	0.029	0.251	0.116
Hands + Speech	208	0.027	0.113	0.243

Looking at the symmetric conditional probability (the product of each of the one-way conditional probabilities between the two behaviors) of the bigrams, we can see which bigram pairs show the greatest dependency. The highest ranked bigram is gaze-towards followed by a head onset, and two more of the top four also involve head onsets and gaze shift. This kind of pattern suggests a larger pattern, and indeed, when we look at 3-grams, we see that the most frequent is gaze-towards + head onset + gaze-away. Also highly ranked are bigrams of speech + head and head + speech, and we see that for speakers, speech onset is somewhat more likely to precede head onset. It is also worth noting that manual gesture onsets tend are more likely to precede head gesture and speech onsets.

Bigrams of listener onsets across modalities are shown in Table 11.

Table 11. Listener onset bigrams by modality (1-second window)

Bigram (1 + 2)	**Frequency**	**Symmetric CP**	**CP: 2|1**	**CP: 1|2**
Head + Speech	263	0.083	0.363	0.228
Head + Gaze-away	106	0.023	0.248	0.092
Head + Gaze-towards	98	0.018	0.215	0.085
Gaze-away + Speech	71	0.016	0.098	0.166
Speech + Head	113	0.015	0.098	0.156
Speech + Gaze-away	66	0.014	0.154	0.091
Gaze-towards + Head	80	0.012	0.069	0.175
Gaze-away + Head	75	0.011	0.065	0.175
Head + Hands	26	0.009	0.413	0.023

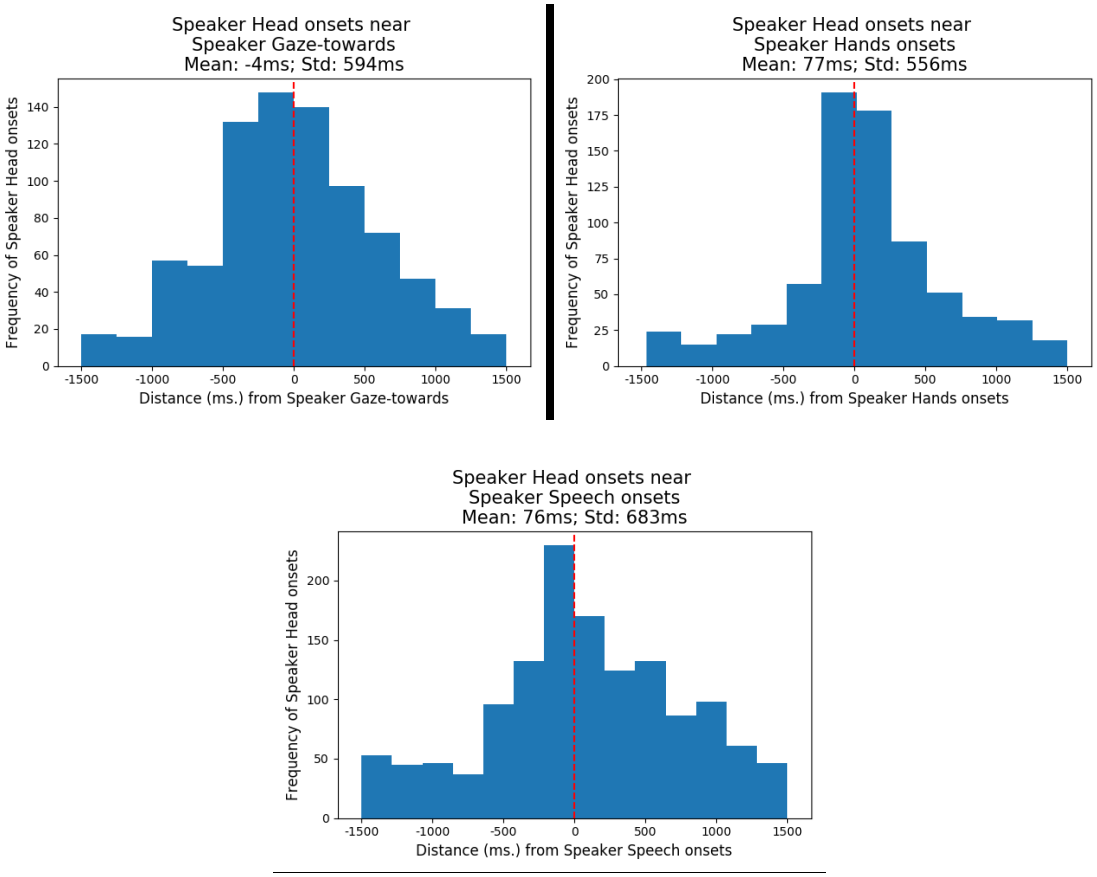
Listeners show several different patterns. For sequences of speech and head onsets, listeners are much more likely to begin to start the head gesture before the speech. Head and gaze onset bigrams are also highly ranked for listeners, but they are more likely to begin a head gesture and then shift their gaze away. The gaze-towards + head gesture onset + gaze-away construction is not common among listeners, but one of their most common 3-grams shows pattern that is almost the reverse: gaze-away + speech onset + gaze-towards. Looking away while speaking seems to be common across both roles.

2.3 Window Histograms

2.3.1 Speaker Onsets near Speaker Onsets

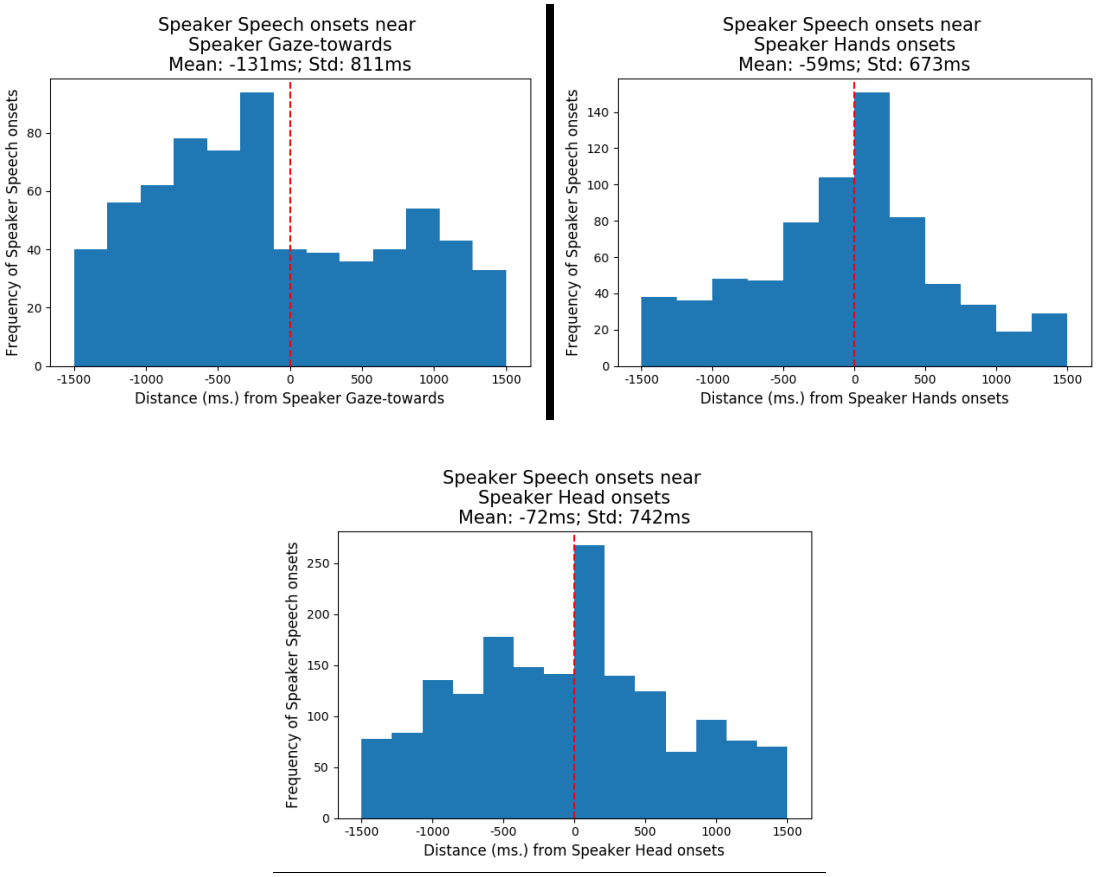
We will now take a closer look at the fine-grained temporal distribution of the behavior boundaries that co-occur near each other. The figures below illustrate how looking at all the timing relations across all behaviors can give a picture of the temporal characteristics of a system.

Figure 29. Window histogram – Speaker head onsets near onsets of other modalities



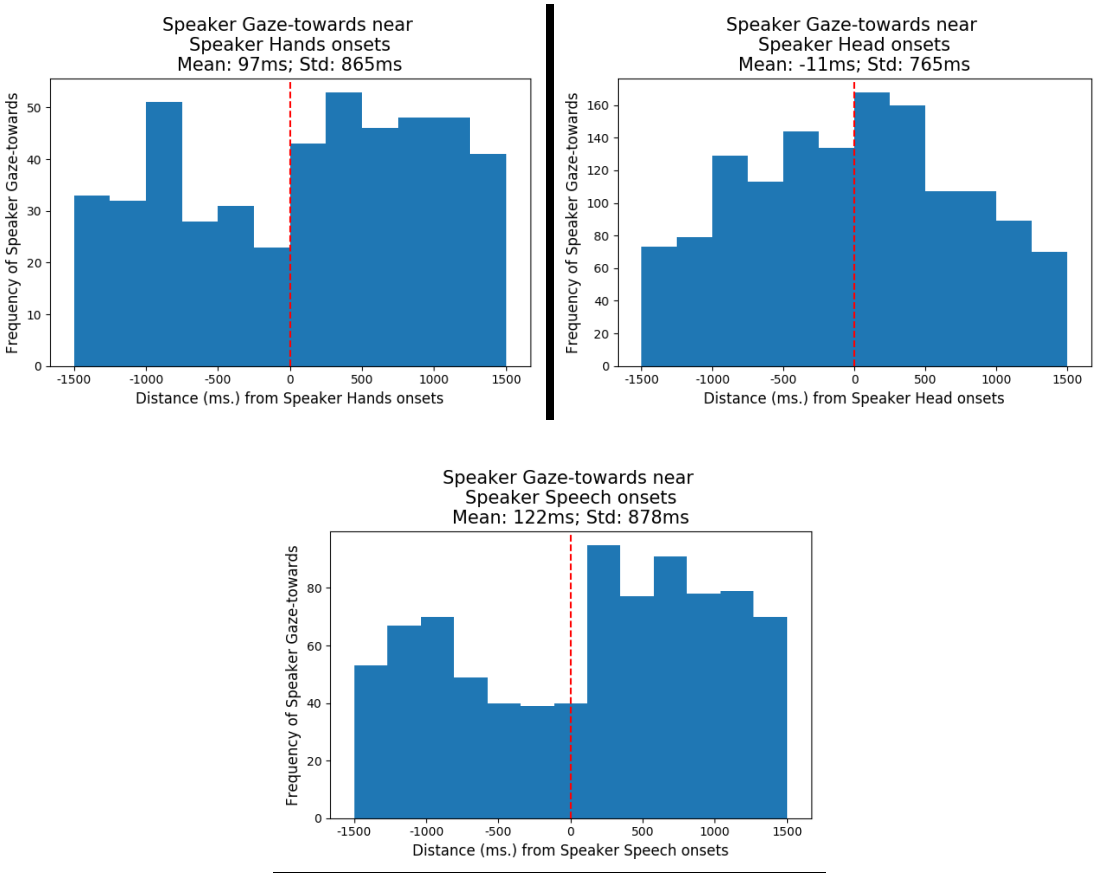
Speaker head onsets tend to be timed to occur near each other modality's onsets, but most precisely with hand onsets.

Figure 30. Window histogram – Speaker speech onsets near onsets of other modalities



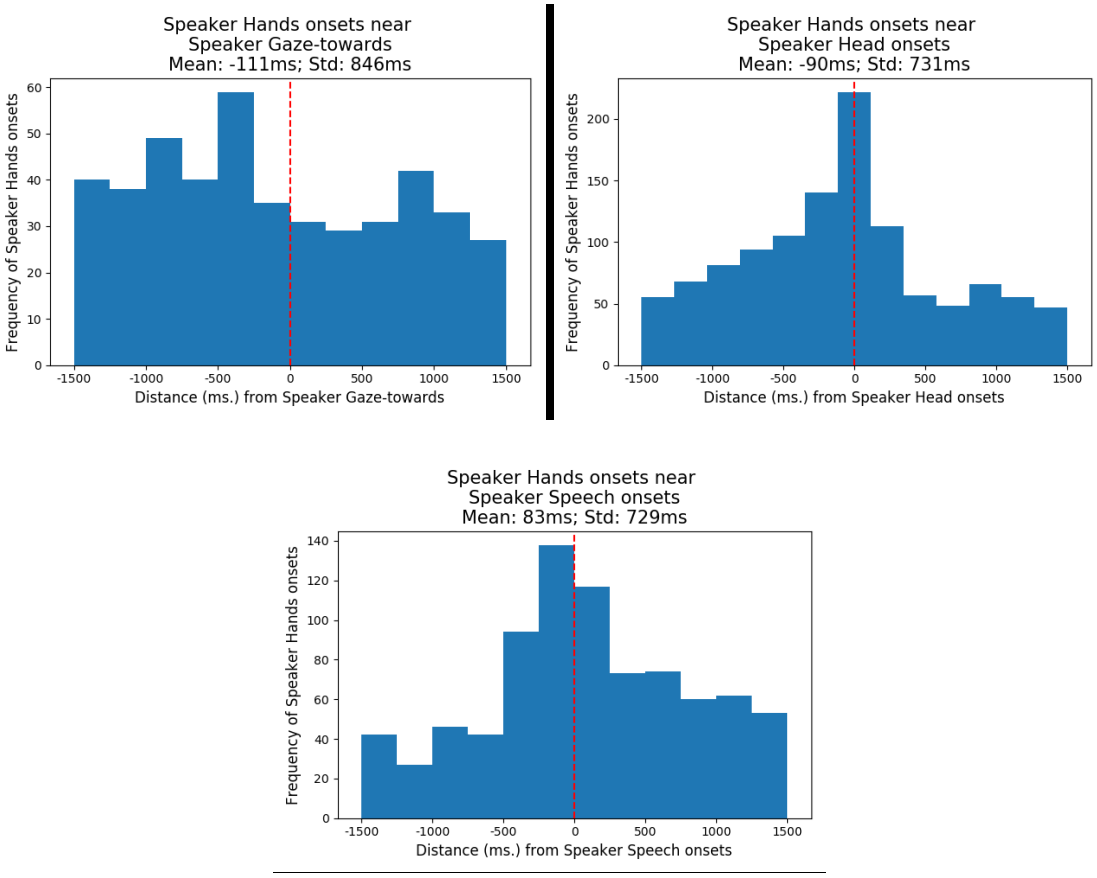
Speaker speech onsets are more likely to precede gaze-shift towards the listener – we see a steep drop-off just before the shift. They also tend to occur near manual gesture onsets, though we can see from this distribution how they peak just after.

Figure 31. Window histogram – Speaker gaze-towards near onsets of other modalities



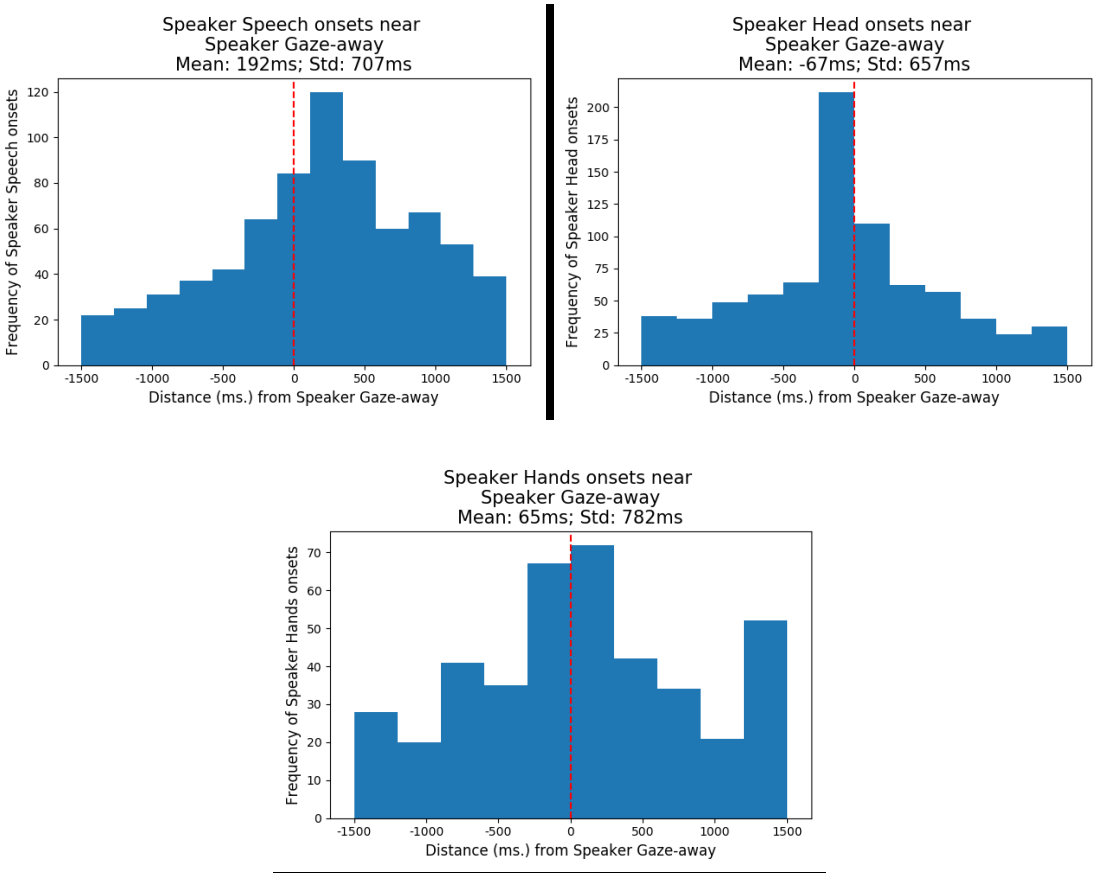
Looking at gaze-shifts towards the listener, we see that these shifts tend to occur more after speech and manual gesture onsets.

Figure 32. Window histogram – Speaker manual gesture onsets near onsets of other modalities



We see the same well-timed peaks between manual gesture onsets and head and speech onsets. We don't see the same degree of more frequent hand onsets before gaze-towards as we saw earlier, but we should remember that these are not precisely the same distributions (that is, this figure selects all manual gesture onsets within 1500ms of gaze-towards onsets, while the former selected all gaze-towards onsets within 1500ms of manual gesture onsets – only the 1500ms between these boundaries is shared across the two figures, not the 1500ms on either side).

Figure 33. Window histogram – Speaker onsets of other modalities near Speaker gaze-away

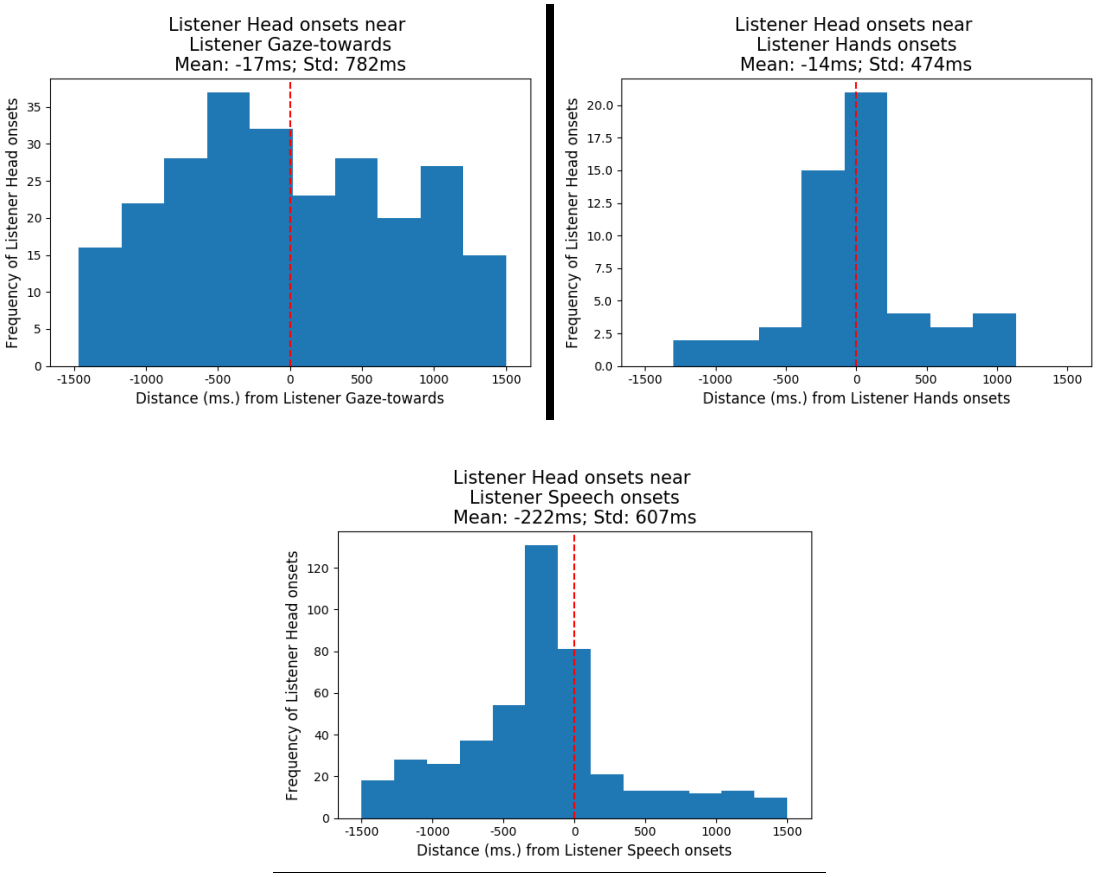


We can also look at how these onsets are timed with speaker gaze-away. Gaze-away sometimes occurs at the moment of the onset of communicative behavior: speech peaks just after and then diminishes, head gestures peak just before, and manual gesture peaks concurrently.

2.3.2 Listener Onsets Near Listener Onsets

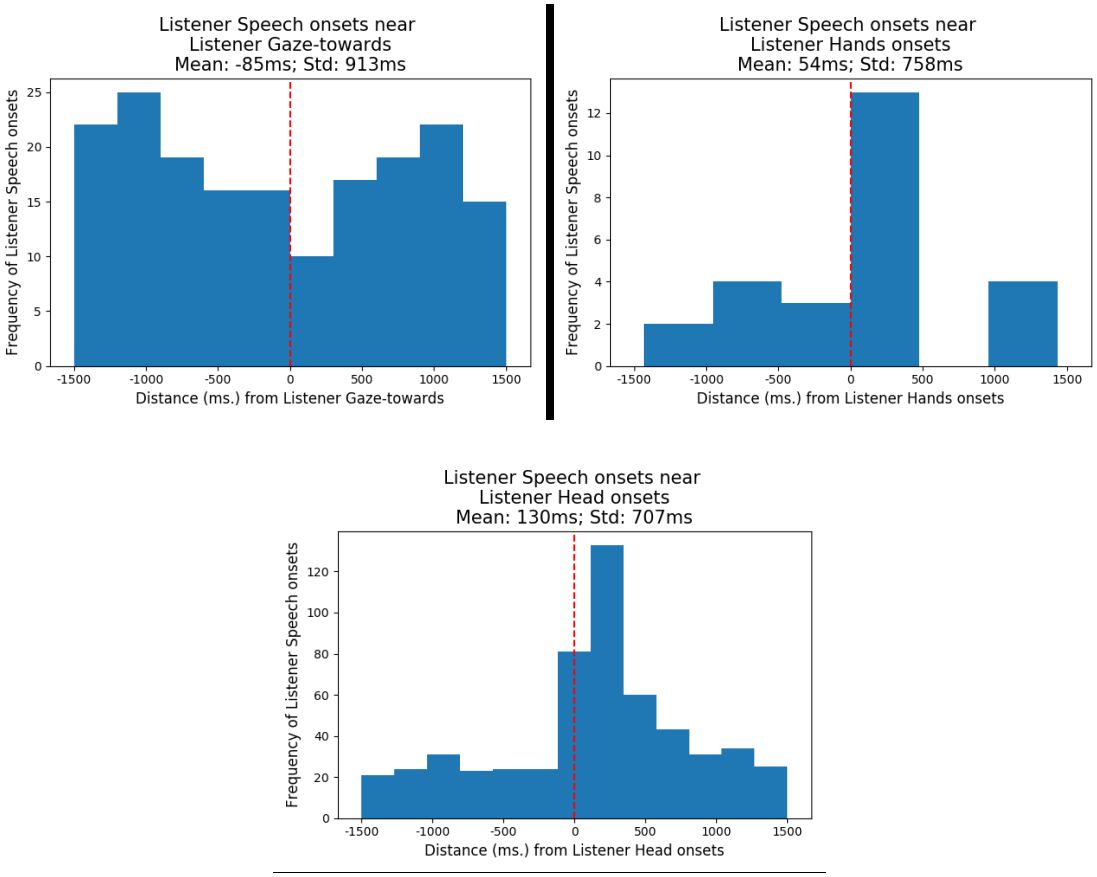
There are fewer instances of all listener behaviors, particularly listener manual gesture, so the distributions in the following figures may not show results as robustly as those for speakers.

Figure 34. Window histograms – Listener head onsets near onsets of other modalities



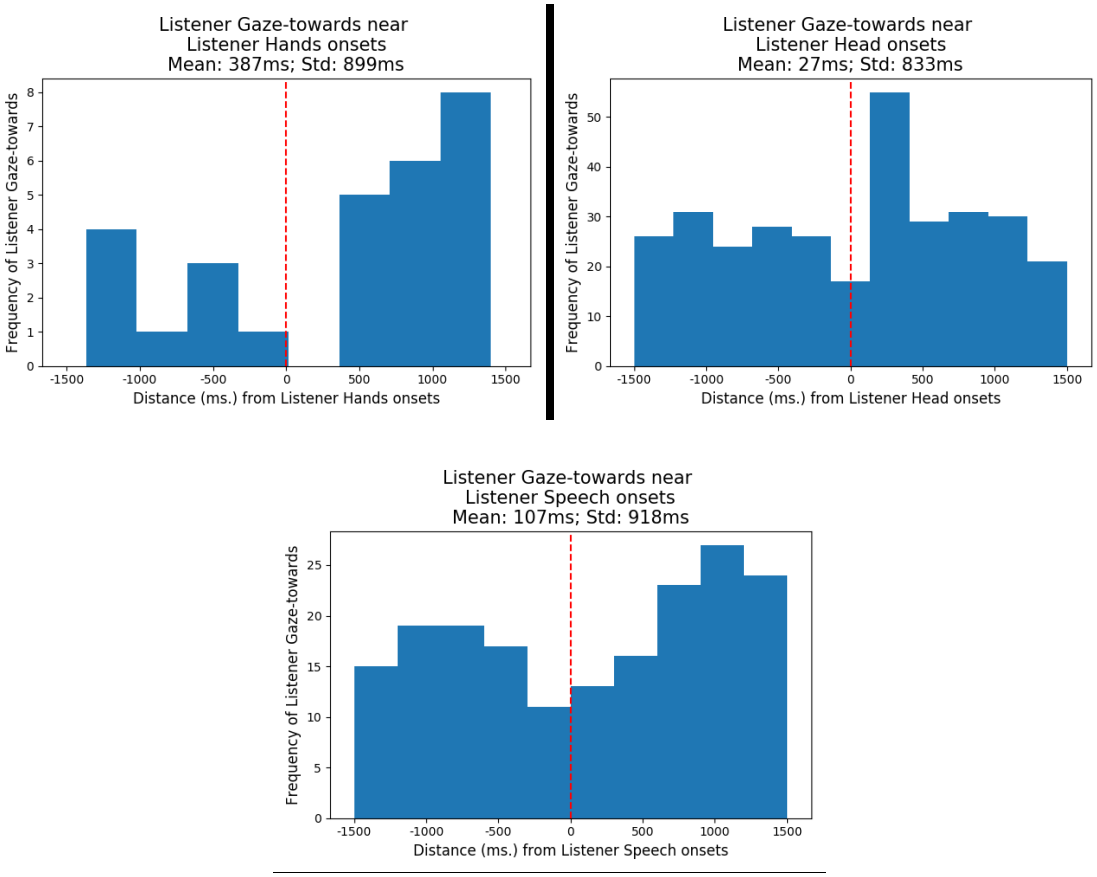
While listeners do not seem to time their head onsets with respect to gaze-shift towards the speaker, they very clearly time them to correspond with manual gesture and speech onsets. We can see very clearly how they precede speech onsets, and how they are more likely to slightly follow manual gesture onsets.

Figure 35. Window histograms – Listener speech onsets near onsets of other modalities



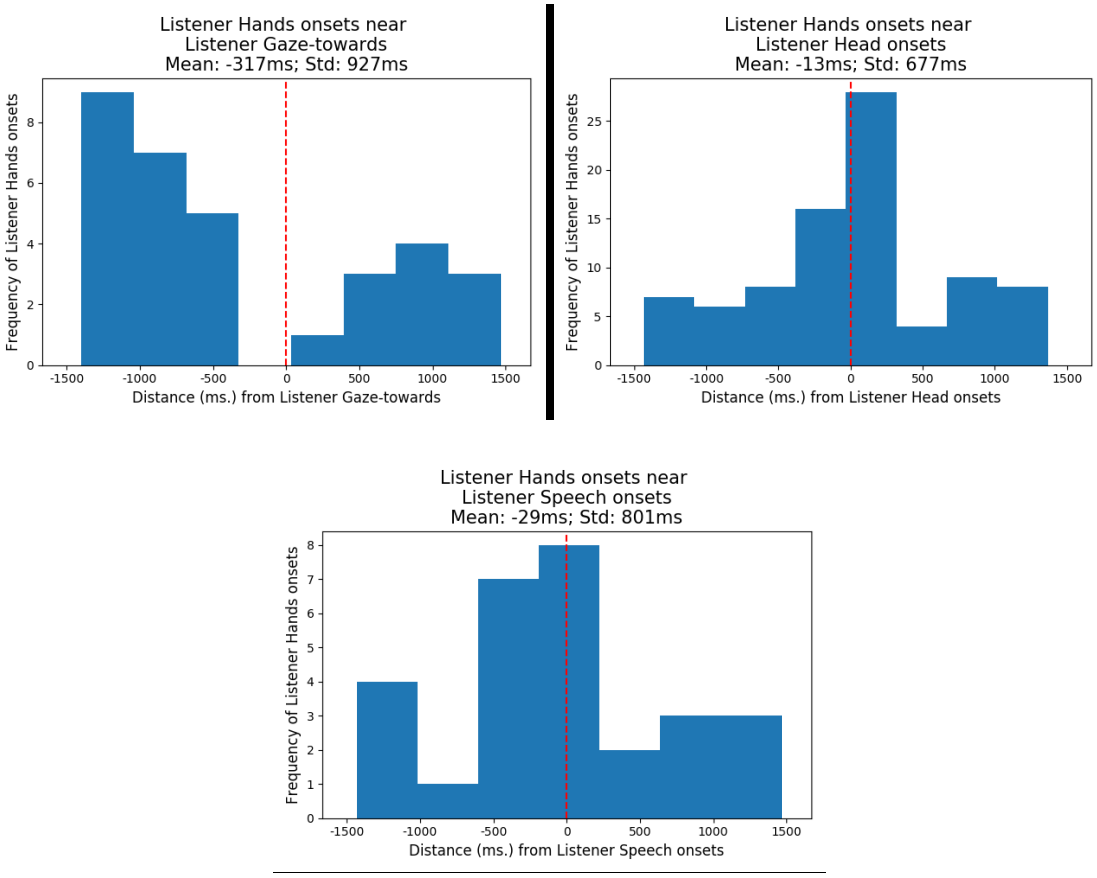
Listener speech onsets, as we saw in Table 10, tend to follow manual and head gesture onsets. We see almost the opposite of a peak with gaze-towards, with speech onsets being more likely to occur before or after, but not simultaneously with, a gaze-shift towards.

Figure 36. Window histogram – Listener gaze-towards near onsets of other modalities



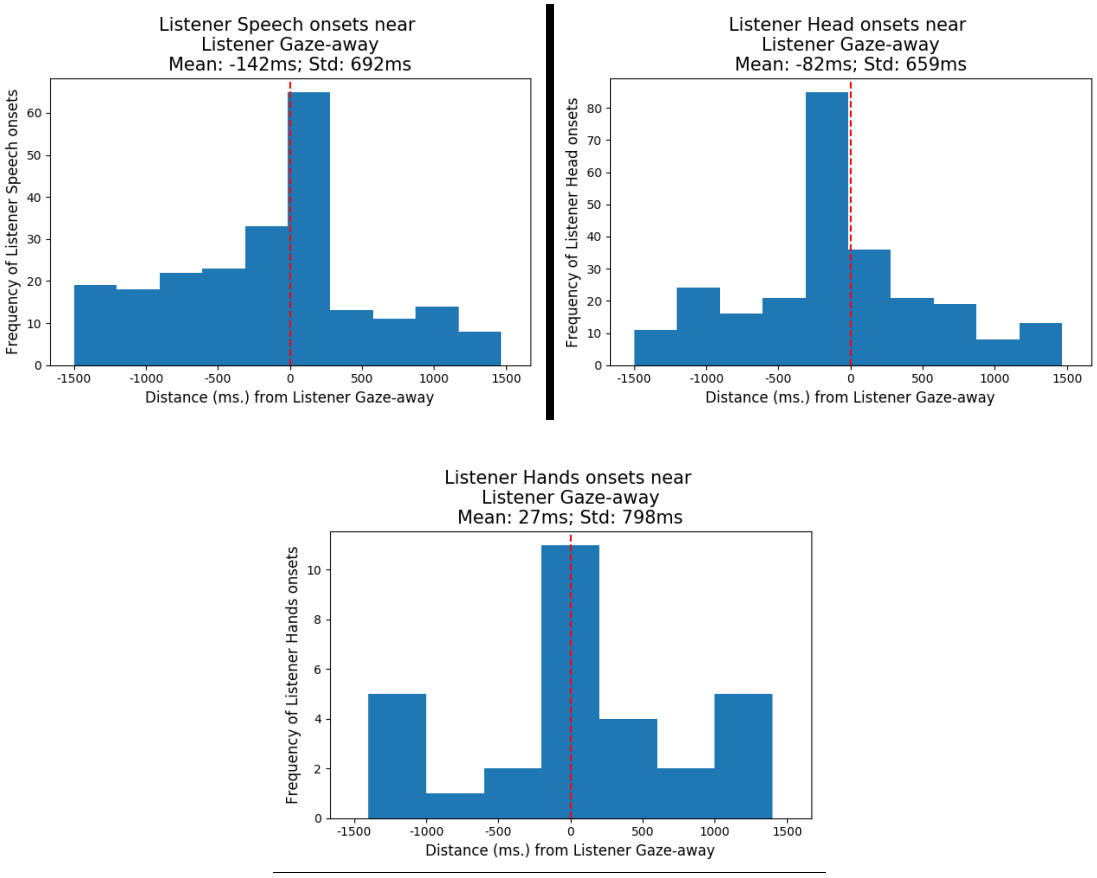
There is a tendency for listener gaze-towards to follow manual gesture – presumably when listeners have just looked away to speak and co-gesture. There is a small peak in gaze-shift directly after a head onset.

Figure 37. Window histograms – Listener manual gesture onsets near onsets of other modalities



We see here what we have seen in Figures 34 and 35, with listener manual gesture onsets tending to co-occur with head gesture and speech onsets, and tending to follow more than precede listener gaze-towards.

Figure 38. Window histogram – Listener onsets of other modalities near Listener gaze-away



Listener behaviors seem to be much more precisely timed to occur with gaze-shift away than with gaze-shift towards (as was the case for speakers, as well). And, like with speakers, listener speech tends to immediately follow gaze-away, and listener head onsets tend to occur immediately before. Manual gesture occurs simultaneously.

3. Heads and Speech

We now turn to the timing relations between subtypes of head and speech behavior. These two modalities have the most detailed coding schemes, and so the number of possible interactions between coded behaviors is the largest.

3.1 Speakers

3.1.1 Likelihood Measures

To start with, we'll look at the overlap between speaker speech and each speaker head subtype (Table 12). The one-way conditional probabilities of this overlap are shown next to the odds ratio of the overlap.

Table 12. Conditional probabilities and odds ratios of Speaker Speech (all) and Speaker Head Subtypes

	CP|Speech	CP|Head	Odds ratio
Single wag	0.006	0.920	**4.077*** **
Jut in	0.015	0.920	**4.103*** **
Multiple shake	0.066	0.878	**2.674*** **
Tilt away + return	0.006	0.875	**2.470*** **
Multiple wag	0.010	0.858	**2.138*** **
Single jut	0.014	0.856	**2.124*** **
Nod down	0.023	0.826	**1.699*** **
Single shake	0.017	0.826	**1.689*** **
Single retraction	0.009	0.826	**1.678*** **
Single nod	0.040	0.801	**1.447*** **
Tilt towards	0.014	0.795	**1.373*** **
Tilt away	0.012	0.794	**1.363*** **
Nod up	0.014	0.748	ns.
Tilt towards + return	0.007	0.742	ns.
Retraction back	0.007	0.738	ns.
Multiple nod	0.042	0.687	0.760***

We saw in Figure 26 that speaker speech and head gesture overlap more than expected, given a random distribution of the behaviors with the same frequencies, and here we can see how that dependency is broken down for different types of head gesture. It seems

that, of all the subtypes, multiple nods are actually less likely to overlap than expected. Also, many half-cycle gestures, with the exception of juts-in, are not strongly attracted to speech.

Looking the opposite direction, Table 13 (below) shows the dependencies between speaker heads and different speaker speech types.

Table 13. Conditional probabilities and Odds ratios of Speaker Head with Speaker Speech Subtypes

	CP|Speech	**CP|Head**	**Odds Ratios**
Affirmation	0.537	0.012	**2.899*****
Laugh	0.204	0.002	0.634**
Back-channel	0.561	0.018	**3.199*****
Declarative	0.338	0.603	**1.666*****
Interrogative	0.271	0.100	0.913*
Incomplete	0.255	0.081	0.833*
Filler	0.093	0.009	0.249***

There are only two speaker back-channel types with enough tokens to warrant examination. Heads overlap with over fifty percent of affirmations (almost three times more likely than expected), but overlap with laughs is less than expected.

Head gesture makes up a greater percentage of back-channels than any subtype of speech turn, but is most strongly associated with declaratives, and least associated with fillers. Correspondingly, declaratives overlap with over 60% of head gestures, while fillers overlap with less than 1% of them.

The following tables look in detail at the overlaps between head and speech subtypes.

Table 14. Conditional Probabilities and odds ratios of Speaker head behaviors with Speaker declarative and interrogative speech

	Declarative Cond. Prob.	Declarative Odds-ratio			Interrogative Cond. Prob.	Interrogative Odds-ratio
Single wag	0.721	**2.546*****		**Tilt away + return**	0.209	**2.236*****
Multiple shake	0.718	**2.464*****		**Single retraction**	0.190	**1.994*****
Multiple wag	0.706	**2.291*****		**Jut in**	0.188	**1.968*****
Jut in	0.665	**1.900*****		**Tilt away**	0.140	**1.372*****
Single shake	0.660	**1.862*****		**Nod down**	0.137	**1.344*****
Tilt away + return	0.658	**1.832*****		**Single jut**	0.131	**1.270****
Tilt towards	0.646	**1.748*****		**Multiple wag**	0.118	ns.
Single jut	0.628	**1.614*****		**Nod up**	0.115	ns.
Single nod	0.595	**1.416*****		**Single nod**	0.105	ns.
Nod down	0.586	**1.355*****		**Multiple nod**	0.087	0.788**
Tilt away	0.574	**1.283*****		**Tilt towards + return**	0.074	0.668***
Single retraction	0.565	**1.233****		**Tilt towards**	0.073	0.656***
Tilt towards + return	0.548	**1.151***		**Multiple shake**	0.072	0.642***
Retraction back	0.537	**1.082***		**Single wag**	0.069	0.626***
Multiple nod	0.531	ns.		**Single shake**	0.066	0.591***
Nod up	0.485	0.890**		**Retraction back**	0.065	0.580***

With speaker declaratives, most head gesture subtypes are more likely than expected. This attraction is strongest for wags, shakes, and gestures with motion towards the listener (juts and tilts ending with forward motion). It is weakest with nods-up, multiple nods, and gestures ending in motion away from the listener.

Interrogatives show more diversity in attraction and repellence. Here, four of the six gestures that are more likely than expected involve motion towards the listener, while two of the gestures that are less likely than expected involve motion away (tilt towards +

return and retraction back). However, the half-cycle tilts seem to break this pattern, with tilts-towards being repelled and tilts-away being retracted. Also less likely than expected are single and multiple head shakes.

The single retraction patterns with juts here, and this led to closer investigation of its kinematic characteristics. The single retraction gesture involves motion away from the listener followed by motion towards. It was found that this motion towards tended to be the faster, more salient part of the gesture, as though the head is being retracted to prepare for the forward motion (similar to the tilt away + return). The reverse is not true for single juts – for these, the motion forward is also the more salient part of the gesture, giving us a categorical asymmetry.

Table 15 describes the overlaps between speaker head gestures and fillers and incompletes.

Table 15. Conditional Probabilities and odds ratios of Speaker head behaviors with Speaker filler and incomplete speech

	Filler Cond. Prob.	Filler Odds-ratio			Incomplete Cond. Prob.	Incomplete Odds-ratio
Retraction back	0.027	ns.		**Nod up**	0.131	**1.524*****
Nod up	0.018	0.644**		**Single wag**	0.129	**1.489*****
Single jut	0.015	0.528***		**Tilt towards + return**	0.110	**1.239****
Tilt away	0.014	0.523***		**Retraction back**	0.105	**1.177***
Nod down	0.014	0.495***		**Single shake**	0.093	ns.
Tilt towards + return	0.010	0.363***		**Nod down**	0.088	ns.
Single nod	0.009	0.332***		**Single nod**	0.088	ns.
Multiple wag	0.007	0.237***		**Single jut**	0.083	ns.
Single shake	0.007	0.233***		**Multiple shake**	0.083	0.895*
Multiple shake	0.005	0.172***		**Tilt away**	0.076	0.825*
Multiple nod	0.004	0.150***		**Tilt towards**	0.072	0.769**
Tilt towards	0.004	0.150***		**Single retraction**	0.071	0.759*
Jut in	0.004	0.130***		**Jut in**	0.067	0.715**
Tilt away + return	0.000	0.000***		**Multiple nod**	0.060	0.618***
Single retraction	0.000	0.000***		**Multiple wag**	0.047	0.493***
Single wag	0.000	0.000***		**Tilt away + return**	0.008	0.077***

For fillers, we see that no head gestures are more likely to overlap than expected. However, there is an interesting pattern among the gestures that are *least* unlikely. Except for juts-in, all half-cycle head gestures are among the highest ranked head gestures. These half-cycle gestures, which reposition the head, seem to indicate some sort of shift of perspective, or precede some shift in the narrative, and these may co-occur with filled pauses because filled pauses this is where such shifts in the upcoming speech are being

processed (Goldman-Eisler 1968, Schegloff, Jefferson, & Sacks 1977)[23]. Much less likely than expected are behaviors with forward motion (except the single jut).

For incomplete speech segments, where the speaker has not completed the predication of the clause, there is no clear pattern.

Table 16. Conditional Probabilities and odds ratios of Speaker head behaviors with Speaker back-channel and non-speech

	Back-channel CP	Back-channel Odds-ratio		Non-speech Cond. Prob.	Non-speech Odds-ratio
Multiple nod	0.055	**9.359*****	**Multiple nod**	0.313	**1.315***
Tilt towards	0.020	**2.240*****	**Retraction back**	0.262	ns.
Single nod	0.014	**1.555***	**Tilt towards + return**	0.258	ns.
Tilt away	0.012	ns.	**Nod up**	0.252	ns.
Retraction back	0.010	ns.	**Tilt away**	0.206	0.734*
Single shake	0.008	ns.	**Tilt towards**	0.205	0.728*
Jut in	0.008	ns.	**Single nod**	0.199	0.691**
Tilt towards + return	0.008	ns.	**Single retraction**	0.174	0.596***
Single retraction	0.006	ns.	**Single shake**	0.174	0.592***
Multiple shake	0.004	0.400***	**Nod down**	0.174	0.589***
Nod down	0.004	0.390**	**Single jut**	0.144	0.471***
Nod up	0.003	0.309***	**Multiple wag**	0.142	0.468***
Single wag	0.000	0.000***	**Tilt away + return**	0.125	0.405***
Single jut	0.000	0.000***	**Multiple shake**	0.122	0.374***
Multiple wag	0.000	0.000***	**Jut in**	0.080	0.245***
Tilt away + return	0.000	0.000***	**Single wag**	0.080	0.244***

With back-channels, multiple nods are the clear favorite – multiple nods seem to be the default back-channel head gesture for both roles. Single nods are also likely, and the tilt-

[23] Although fillers may have other functions, such as holding the floor.

towards shares some kinematic features of the nod (when at an angle, it is sort of a sideways nod).

For non-speech, which is all frames in which no (speaker) speech is occurring, only multiple nods are more likely than expected. This is probably both because it is a common back-channel behavior (and general acknowledgment of an interlocutor behavior), and because it is a thinking behavior, or a behavior one does when one doesn't quite know what else to do. Two motion-away gestures (and nods up) are at chance. Shakes, wags, and motion-towards gestures are less likely than expected.

3.1.2 N-grams

We'll look now at n-grams sequences of speech and head subtypes, ranked by their symmetric conditional probabilities (Table 17).

Table 17. Speaker Speech and Head onset bigrams (1-second window)

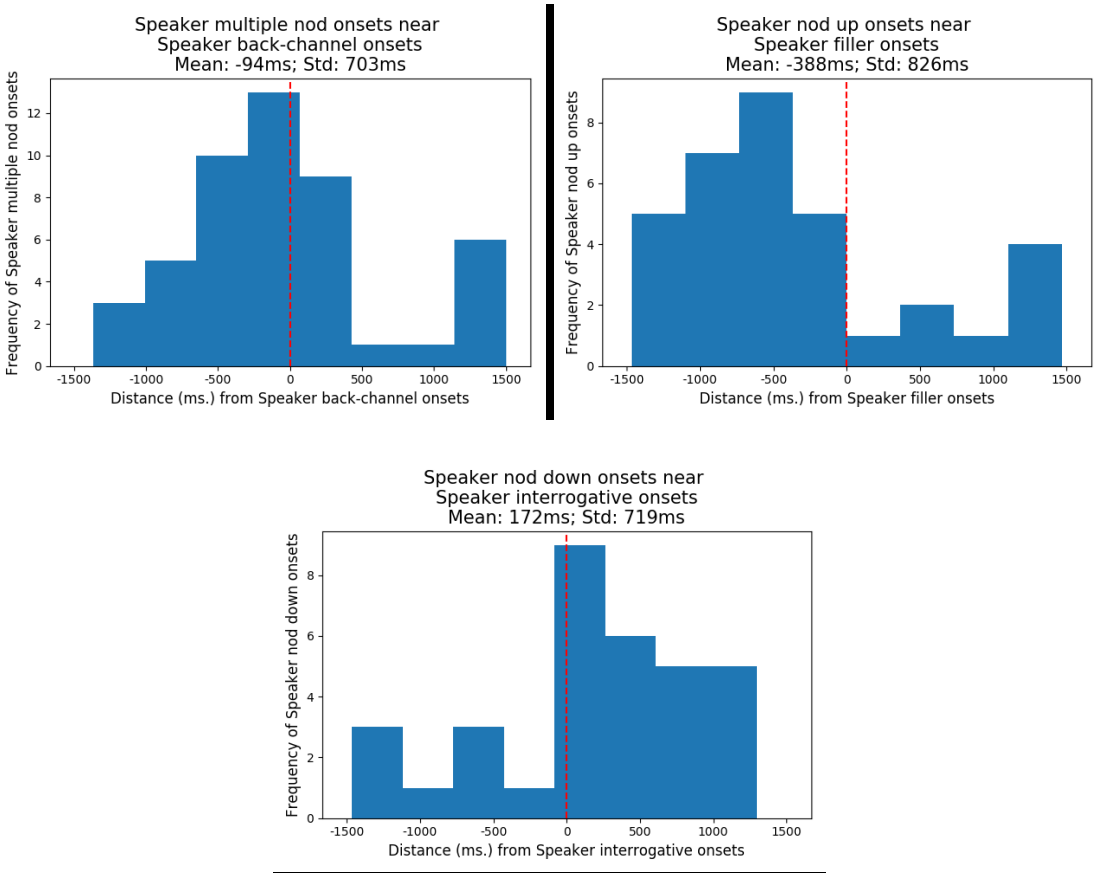
Rank	Bigram	Count	Symm. CP	CP: 2|1	CP: 1|2	Rank	Bigram	Count	Symm. CP	CP: 2|1	CP: 1|2
1	**m. nod + back-ch.**	24	0.022	0.238	0.093	10	**incomplete + nod-down**	20	0.005	0.092	0.059
2	**declarative + m. shake**	54	0.012	0.236	0.052	11	**declarative + s. shake**	26	0.005	0.208	0.025
3	**declarative + s. nod**	58	0.011	0.193	0.056	12	**m. nod + declarative**	37	0.005	0.036	0.143
4	**nod-up + filler**	18	0.008	0.069	0.117	13	**declarative + nod-down**	33	0.005	0.152	0.032
5	**s. nod + declarative**	49	0.008	0.047	0.163	14	**declarative + jut in**	20	0.004	0.217	0.019
6	**back-ch. + m. nod**	13	0.006	0.050	0.129	15	**nod-up + declarative**	23	0.003	0.022	0.149
7	**m. shake + declarative**	38	0.006	0.037	0.166	16	**declarative + nod-up**	23	0.003	0.149	0.022
8	**nod-down + declarative**	36	0.006	0.035	0.166	17	**m. nod + incomplete**	17	0.003	0.050	0.066
9	**interrogative + nod-down**	16	0.006	0.074	0.075	18	**nod-up + incomplete**	13	0.003	0.038	0.084

We see in Table 17 that speaker speech is more likely to precede head gesture than follow it. We do not see any particular patterns leading us to believe that some head types always precede speech types, or vice versa, but we do see which pairs are most common. The beginnings of back-channels are associated with multiple nods, the beginning of declaratives are associated with single nods, nods-down, and multiple shakes, the beginnings of interrogatives are associated with nods-down, and the beginnings of fillers are associated with nods up.

3.1.3 Window Histograms

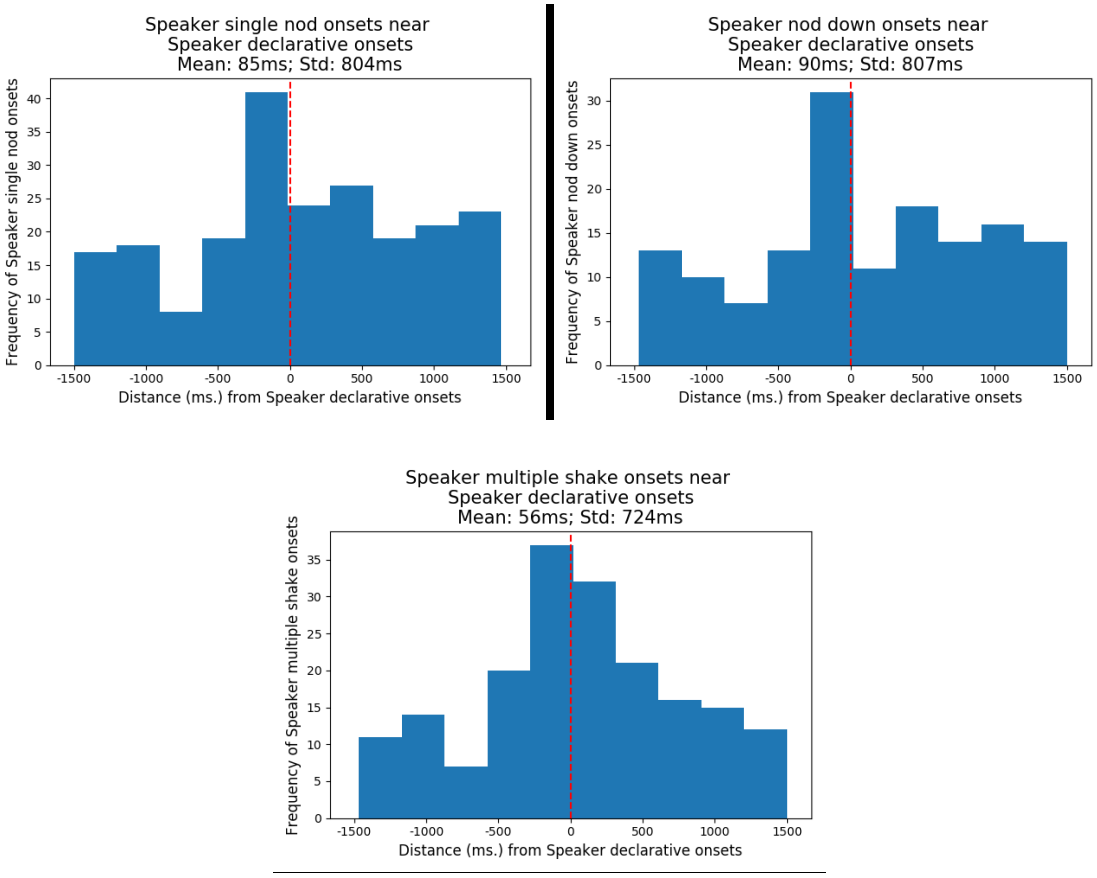
Given the low frequency of some subtypes, not all window histograms will have enough tokens to show interesting results. However, we can look at the bigram patterns from Table 17.

Figure 39. Window histogram – Speaker multiple nod onsets near speech onsets



Multiple nods near back-channels and single nods near fillers both show a strong tendency to precede their accompanying speech onset. This is not the case for nods-down near interrogatives, which tend to come after the speech onset.

Figure 40. Window histograms – Speaker single nod, nod down, and multiple shake onsets near declarative onsets



For the different head gestures associated with declarative onsets, we see small peaks for single nods and nods-down, and a stronger peak for multiple shakes, tapering off in frequency after the declarative onset.

3.2 Listeners

3.2.1 Likelihood Measures

Listeners have far fewer tokens of different speech turns, and far more tokens of back-channels, so back-channels and head gestures will be the focus of this section. First, we'll look at the overlaps between head gestures and speech turn vs. back-channels (Table 18).

Table 18. Conditional Probabilities and odds ratios of Listener head behaviors with Listener speech turns and back-channels

	Speech-turn CP	Speech-turn Odds-ratio			Back-channel CP	Back-channel Odds-ratio
Single shake	0.400	**10.834*****		**Single shake**	0.289	**6.300*****
Multiple shake	0.359	**9.398*****		**Jut in**	0.272	**5.782*****
Tilt towards + return	0.311	**7.329*****		**Tilt towards + return**	0.269	**5.689*****
Single retraction	0.252	**5.439*****		**Retraction back**	0.227	**4.566*****
Jut in	0.243	**5.189*****		**Single jut**	0.213	**4.186*****
Single jut	0.233	**4.900*****		**Single retraction**	0.192	**3.658*****
Nod down	0.184	**3.678*****		**Tilt towards**	0.168	**3.120*****
Retraction back	0.175	**3.415*****		**Nod up**	0.150	**2.733*****
Tilt towards	0.170	**3.313*****		**Single nod**	0.138	**2.545*****
Tilt away	0.167	**3.241*****		**Nod down**	0.131	**2.339*****
Single nod	0.140	**2.711*****		**Multiple nod**	0.116	**2.256*****
Tilt away + return	0.138	**2.572*****		**Tilt away + return**	0.087	**1.468****
Nod up	0.105	**1.892*****		**Tilt away**	0.087	**1.468****
Multiple nod	0.040	0.645***		**Multiple shake**	0.069	ns.

As we saw with speakers, almost all head gestures are more likely than expected to overlap. The exception here is the multiple nod. The most likely are shakes and, as seen with speakers, head gestures involving motion-towards the speaker.

Unlike with speakers, almost all head gesture types are also more likely than expected to co-occur with back-channels. The major differences between speech turns and back-channels are that multiple nods are more likely for listeners, and multiple shakes are only at chance. But aside from these, it seems that listeners utilize almost all forms of head gesture to co-produce back-channels. If it is true that different spoken back-channels have different functions (which seems to be the consensus in back-channel research), then if

different head gestures pattern with different kinds of spoken back-channels, this will suggest a similar diversity of functions for head gestures.

Of the speech turn categories, only declaratives and interrogatives had enough tokens for listener analysis (Table 19).

Table 19. Conditional Probabilities and odds ratios of Listener head behaviors with Listener declarative and interrogative speech

	Declarative CP	**Declarative Odds-ratio**		**Interrogative CP**	**Interrogative Odds-ratio**
Single shake	0.336	**13.967*****	**Single jut**	0.106	**5.629*****
Tilt towards + return	0.305	**12.135*****	**Jut in**	0.087	**4.519*****
Multiple shake	0.301	**12.480*****	**Nod down**	0.072	**3.718*****
Single retraction	0.198	**6.723*****	**Single retraction**	0.054	**2.723*****
Tilt away	0.143	**4.580*****	**Retraction back**	0.053	**2.652*****
Nod down	0.111	**3.418*****	**Single shake**	0.049	**2.415*****
Tilt towards	0.109	**3.333*****	**Tilt towards + return**	0.044	**2.200*****
Single jut	0.103	**3.134*****	**Multiple shake**	0.041	**2.058*****
Single nod	0.092	**2.841*****	**Single nod**	0.041	**2.089*****
Tilt away + return	0.090	**2.672*****	**Nod up**	0.039	**1.905*****
Retraction back	0.082	**2.412*****	**Tilt towards**	0.035	**1.707*****
Jut in	0.058	**1.665****	**Tilt away**	0.016	ns.
Nod up	0.041	ns.	**Multiple nod**	0.011	0.513***
Multiple nod	0.028	0.761***	**Tilt away + return**	0	0***

Shakes are much more likely than expected, which may come from the fact that listeners are often responding sympathetically to events in the speakers' stories. Shakes are much less highly ranked with interrogatives. Tilts away + return never co-occur, while they are

among the most likely to co-occur with listener declaratives – this follows the same pattern as with speakers, but to a much stronger degree.

We now look at overlaps between listener head gestures and listener back-channels, starting with the two most frequent back-channels: acknowledgments and assessments (Table 20).

Table 20. Conditional Probabilities and odds ratios of Listener head behaviors with Listener acknowledgments and assessments

	Acknowledg-ment CP	Acknowledg-ment Odds-ratio		Assessment CP	Assessment Odds-ratio
Tilt towards + return	0.121	**8.804*****	**Retraction back**	0.195	**7.627*****
Single jut	0.091	**6.310*****	**Single retraction**	0.172	**6.472*****
Tilt away + return	0.087	**6.060*****	**Jut in**	0.162	**6.036*****
Multiple nod	0.052	**4.816*****	**Single shake**	0.146	**5.319*****
Jut in	0.049	**3.257*****	**Tilt towards**	0.103	**3.608*****
Single shake	0.046	**3.058*****	**Nod up**	0.081	**2.746*****
Single nod	0.046	**3.177*****	**Single jut**	0.080	**2.687*****
Tilt towards	0.040	**2.636*****	**Nod down**	0.069	**2.294*****
Nod up	0.031	**2.045*****	**Single nod**	0.056	**1.877*****
Nod down	0.031	**1.987*****	**Tilt away**	0.056	**1.826*****
Retraction back	0.027	**1.754****	**Tilt towards + return**	0.053	**1.715****
Single retraction	0.020	ns.	**Multiple nod**	0.016	0.483***
Tilt away	0.013	ns.	**Multiple shake**	0.015	0.459***
Multiple shake	0.009	0.547**	**Tilt away + return**	0.000	0.000***

Although these acknowledgments don't make up a huge percentage of any given head behaviors, most head behaviors are more likely than expected to overlap. Motion-towards behaviors are among the more likely, as are multiple and single nods. Single shakes are

more likely, but multiple shakes are less likely. Away-motion gestures are among the least likely.

For assessments, most head gestures are also more likely than expected to overlap. Here, however, retractions back are the most likely (although motion-towards gestures are also quite likely). We see the same difference between single and multiple shakes, with single shakes being common and multiple shakes being uncommon. Multiple cycle gestures are less likely common here, with even the common multiple nods being less likely than expected.

Table 21. Conditional Probabilities and odds ratios of Listener head behaviors with Listener continuers and affirmations

	Continuer CP	Continuer Odds-ratio		Affirmation CP	Affirmation Odds-ratio
Tilt towards + return	0.044	**7.691*****	**Single shake**	0.060	**16.746*****
Multiple nod	0.024	**5.977*****	**Tilt towards + return**	0.036	**9.830*****
Single nod	0.015	**2.616****	**Multiple shake**	0.020	**5.403*****
Tilt towards	0.014	**2.273***	**Tilt away**	0.019	**4.901*****
Nod down	0.007	ns.	**Single nod**	0.015	**4.330*****
Nod up	0.004	ns.	**Nod down**	0.012	**3.241*****
Single retraction	0.000	ns.	**Multiple nod**	0.012	**4.141****
Jut in	0.000	ns.	**Tilt towards**	0.011	**2.736***
Single shake	0.000	ns.	**Retraction back**	0.000	ns.
Single jut	0.000	ns.	**Single retraction**	0.000	ns.
Tilt away + return	0.000	ns.	**Jut in**	0.000	ns.
Tilt away	0.000	0.000***	**Single jut**	0.000	ns.
Multiple shake	0.000	0.000***	**Tilt away + return**	0.000	ns.
Retraction back	0.000	0.000***	**Nod up**	0.000	0.000**

For continuers, the nod is the dominant behavior, with all single and multiple nods showing greater overlap than expected. Tilts towards and tilts towards + return are the next most likely, which are kinematically similar to nods down and single and multiple nods, in that both involve forward motion – for juts this is the entire head, while for nods it is only the forehead.

For affirmations, interestingly, while all nods (except nods-up, which never overlap) are more likely to overlap than expected, both single and multiple head shakes are even more likely, which might be unexpected in an affirming context (head shakes express sympathy rather than negation). As with continuers, we see tilts towards and tilts towards + return being more likely than expected, again patterning with nods.

Table 22 shows the conditional overlaps between two low-frequency listener back-channels (collaborative finishes and newsmarkers) with head gesture types.

Table 22. Conditional Probabilities and odds ratios of Listener head behaviors with Listener collaborative finishes and newsmarkers

	Coll. Finish CP	Coll. Finish Odds-ratio		Newsmarker CP	Newsmarker Odds-ratio
Single jut	0.043	**13.863*****	**Tilt towards + return**	0.053	**27.033*****
Multiple shake	0.027	**8.967*****	**Nod down**	0.014	**6.746*****
Single shake	0.013	**4.029*****	**Single nod**	0.014	**7.277*****
Multiple nod	0.010	**3.858****	**Nod up**	0.008	**3.777****
Single nod	0.003	ns.	**Multiple nod**	0.006	**3.278****
Tilt towards + return	0.000	ns.	**Single jut**	0.000	ns.
Tilt away	0.000	ns.	**Multiple shake**	0.000	ns.
Tilt towards	0.000	ns.	**Single shake**	0.000	ns.
Nod up	0.000	ns.	**Tilt away**	0.000	ns.
Retraction back	0.000	ns.	**Tilt towards**	0.000	ns.
Single retraction	0.000	ns.	**Retraction back**	0.000	ns.
Jut in	0.000	ns.	**Single retraction**	0.000	ns.
Tilt away + return	0.000	ns.	**Jut in**	0.000	ns.
Nod down	0.000	0.000***	**Tilt away + return**	0.000	ns.

Collaborative finishes and newsmarkers are among the least frequent listener back-channels. Collaborative finishes tend to co-occur with juts, shakes, and multiple nods, while newsmarkers tend to co-occur with nods and tilts towards + return.

Table 23. Conditional Probabilities and odds ratios of Listener head behaviors with Listener laughs and non-speech

	Laugh CP	Laugh Odds-ratio		Non-speech CP	Non-speech Odds-ratio
Tilt towards	0.133	**11.224*****	**Multiple shake**	0.912	0.878*
Tilt away + return	0.090	**6.966*****	**Multiple nod**	0.874	0.540**
Tilt away	0.062	**4.711*****	**Nod down**	0.858	0.506*
Single nod	0.040	**3.052****	**Tilt away**	0.851	0.478**
Single jut	0.026	**1.856***	**Nod up**	0.850	0.475*
Nod up	0.025	**1.810***	**Tilt away + return**	0.823	0.391***
Retraction back	0.013	ns.	**Single nod**	0.811	0.348***
Nod down	0.010	ns.	**Single retraction**	0.808	0.354***
Tilt towards + return	0.008	ns.	**Jut in**	0.788	0.313***
Multiple nod	0.006	0.384**	**Retraction back**	0.765	0.273***
Single shake	0.000	0.000***	**Single jut**	0.761	0.267***
Single retraction	0.000	0.000***	**Single shake**	0.711	0.206***
Multiple shake	0.000	0.000***	**Tilt towards**	0.699	0.193***
Jut in	0.000	0.000***	**Tilt towards + return**	0.685	0.182***

Finally, we look at laughs and non-speech segments. The most likely head gestures to co-occur with laughs are tilts, and nods up are also common, the kinds of head gestures one might imagine when someone throws their head back and laughs. Laughs and non-speech segments do not attract shakes, multiple cycles, or motion-away gestures. All head gesture is less likely than expected during non-speech, but multiple nods and multiple shakes, as well as several repositioning head gestures, seem to be the least unlikely to occur apart from speech. Gestures with motion towards or away from the speakers are among the least likely.

3.2.2 N-grams

We'll look now at sequences of listener head and back-channel behaviors (Table 24).

Table 24. Listener Head and Speech bigrams (1-second window)

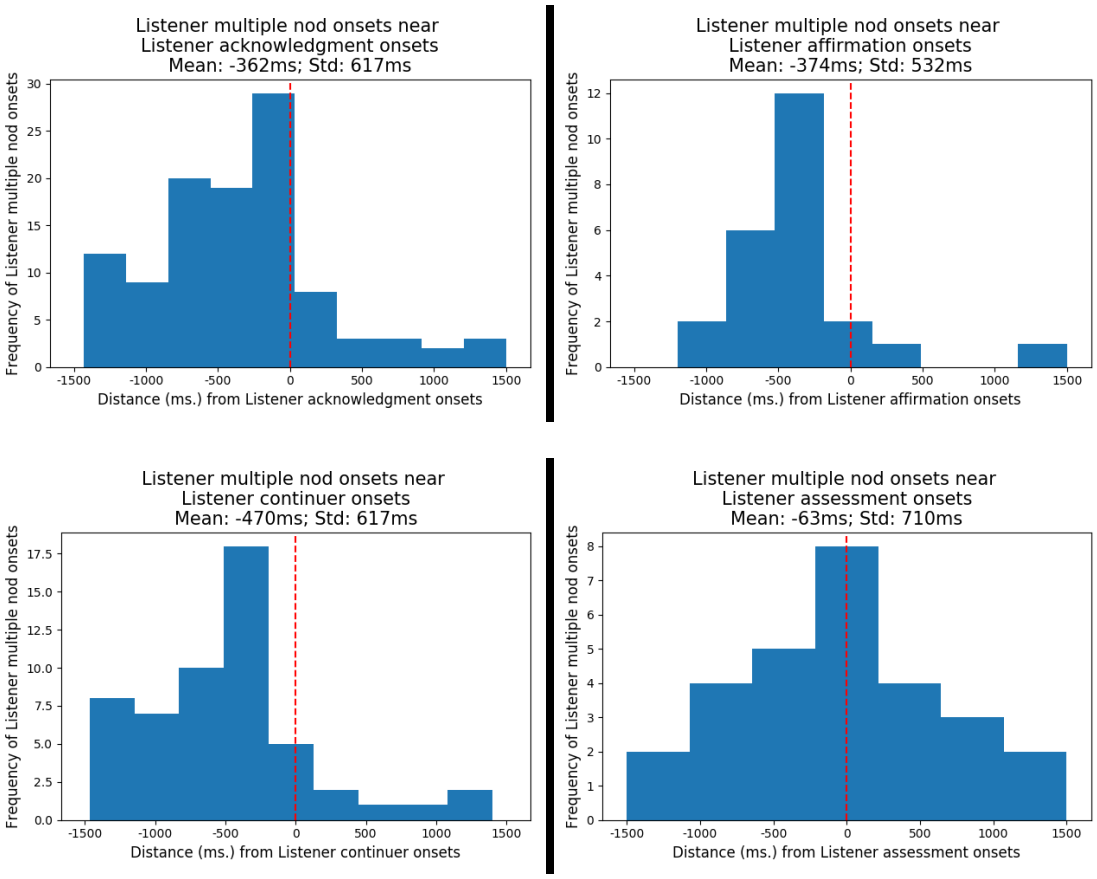
Bigram	**Frequency**	**Symmetric CP**	**CP: 2|1**	**CP: 1|2**
m. nod + acknowl.	69	0.045	0.337	0.133
m. nod + continuer	36	0.026	0.379	0.069
m. nod + affirm.	19	0.017	0.475	0.037
s. nod + acknowl.	24	0.014	0.117	0.117
retract back. + assess.	8	0.010	0.036	0.286
nod-down + newsmark.	3	0.006	0.143	0.045
affirm. + nod-down	4	0.006	0.061	0.100
s. nod + newsmark.	5	0.006	0.238	0.024
acknowl. + nod-down	7	0.004	0.106	0.034
s. nod + continuer	8	0.003	0.084	0.039
s. nod + laugh	9	0.003	0.074	0.044
s. nod + affirm.	5	0.003	0.125	0.024
nod-up + assess.	7	0.003	0.032	0.091

Unlike with speakers, with listeners we can clearly see the tendency of head gesture onsets to precede speech onsets. Multiple nods preceding back-channels are, of course, the most common. They most often precede acknowledgments, continuers, and affirmations. Noticeably absent from these are assessments, which are one of the most frequent listener back-channels, nearly as frequent as acknowledgments. Instead, assessments are more likely to follow retractions of the head, a head gesture often associated with surprise or negative affect, which are common expressions in the assessments in this corpus.

3.2.3 Window Histograms

We can look at the temporal frequency distribution of these multiple nods, relative to the onsets of their accompanying speech.

Figure 41. Window histograms – Listener multiple nod onsets near back-channel onsets



In fact, there is no great difference between the distributions, except that multiple nod onsets near assessments can occur before or after, or that before affirmations and continuers they are likely to peak slightly earlier than before acknowledgments. It's unclear what accounts for these differences, but assessments are the back-channel type that is most similar to a speech turn, and this may help explain why this particular distribution differs from other listener back-channel distributions, looking more like the

interaction between *speaker* head and speech (in Figure 29) than listener head and speech (in Figure 34).

4. Head + Gaze

4.1 Likelihood Measures

Another area where we might expect interactions is between head and gaze. If head gestures are used communicatively in the expectation of a response, we might expect to see more than expected overlap with the head gesturers' gaze, so that the other's response could be seen. We also know that listeners are looking at speakers much more often than the reverse, and so we might expect that speakers will time their head gestures to occur near their gaze-towards more frequently, so as to be able to see whether their head gestures were detected. Table 25 looks at likelihood measures for speaker head behaviors occurring while gazing towards listeners.

Table 25. Conditional Probabilities and odds ratios of Speaker head behaviors with Speaker gaze-towards

	Cond. Prob. / Gaze-towards	Cond. Prob. / Head	Gaze-towards Odds-ratio
Jut in	0.021	0.529	**2.426*****
Multiple nod	0.108	0.517	**2.441*****
Single retraction	0.013	0.445	**1.724*****
Multiple wag	0.013	0.441	**1.695*****
Retraction back	0.013	0.419	**1.548*****
Multiple shake	0.082	0.416	**1.569*****
Tilt towards + return	0.012	0.396	**1.405***
Single nod	0.057	0.393	**1.405****
Single jut	0.017	0.390	**1.373***
Tilt away + return	0.007	0.379	**1.308***
Single wag	0.005	0.345	ns.
Tilt towards	0.018	0.342	**1.114***
Tilt away	0.015	0.334	ns.
Single shake	0.019	0.325	ns.
Nod down	0.022	0.279	0.822*
Nod up	0.012	0.217	0.588**

Speakers do indeed seem to time many of their head gestures to co-occur with gaze-towards at a greater than chance likelihood (or, to look at it another way, it may be that the convention of listener head gestures is that they are made while facing the interlocutor, and to do so while looking away is anomalous, and might be interpreted as being directed at another entity). Several gestures involving motion towards or away are more likely than expected, as are both multiple-cycle gestures. Only the repositioning nods-down and nods-up are less likely than expected.

Table 26. Conditional probabilities and odds ratios of Listener head behaviors with Listener gaze-towards

	Cond. Prob. / Gaze-towards	Cond. Prob. / Head	Gaze-towards Odds-ratio
Multiple nod	0.114	0.901	**1.698****
Single retraction	0.002	0.868	ns.
Tilt away + return	0.003	0.855	ns.
Tilt towards + return	0.003	0.840	ns.
Retraction back	0.004	0.793	0.676**
Tilt towards	0.005	0.754	0.542**
Single jut	0.002	0.746	0.519***
Single nod	0.022	0.743	0.502***
Tilt away	0.005	0.725	0.465***
Jut in	0.002	0.722	0.459***
Nod up	0.005	0.688	0.388***
Single shake	0.002	0.616	0.283***
Multiple shake	0.005	0.447	0.140***
Nod down	0.003	0.373	0.000***

Listeners, on the other hand, do not show the same pattern. Except for multiple nods, all listener head gestures are at chance or less likely to overlap with gaze-towards than expected. Although note the conditional probabilities – these overlaps account for more than half of almost all listener head behaviors, but this is only true for two speaker head behaviors. Single retractions and both full-cycle tilts are at chance. Other behaviors are less likely than expected – all other repositioning gestures, all shakes, most linear z-axis motion, and nods down lead the pack. Nods down and multiple shakes are the two head gestures that happen more when looking away than when looking towards. For listeners, it is fairly common to shake one's head and gaze away, and nods-down often seem to co-occur with processing speech content (see Chapter 7, Section 4.1).

4.2 N-grams

Bigrams of speaker gaze and head behaviors are shown in Table 27.

Table 27. Speaker Gaze and Head bigrams (1-second window)

Bigram	**Frequency**	**Symmetric CP**	**CP: 2|1**	**CP: 1|2**
s. nod + gaze-towards	69	0.016	0.070	0.231
s. nod + gaze-away	65	0.015	0.068	0.217
gaze-towards + s. nod	65	0.014	0.217	0.066
m. nod + gaze-towards	54	0.011	0.055	0.206
gaze-towards + m. nod	49	0.009	0.187	0.050
jut in + gaze-towards	29	0.009	0.030	0.315
m. nod + gaze-away	46	0.008	0.048	0.176
m. shake + gaze-towards	43	0.008	0.044	0.191
gaze-away + nod-up	33	0.007	0.214	0.034
gaze-towards + m. shake	39	0.007	0.173	0.040
m. shake + gaze-away	38	0.007	0.040	0.169
gaze-towards + nod-down	37	0.006	0.171	0.038
nod-up + gaze-towards	31	0.006	0.032	0.201
nod-down + gaze-away	36	0.006	0.037	0.167
nod-up + gaze-away	29	0.006	0.030	0.188

The most frequent kinds of bigrams all involve nodding and gaze-shift, with nods often preceding gaze-towards or gaze-away, and following gaze-towards. These kinds of bigram patterns are often suggestive of longer n-gram patterns in the data, so we take a look at trigram patterns in Table 28.

Table 28. Most frequent Speaker Head and Gaze trigrams (1-second window)

Trigram	Frequency	Trigram	Frequency
gaze-towards + s. nod + gaze-away	27	**gaze-towards + s. shake + gaze-away**	7
gaze-towards + m. nod + gaze-away	12	*gaze-away + jut in + gaze-towards*	6
gaze-towards + nod-down + gaze-away	11	s. nod + gaze-away + gaze-towards	6
gaze-towards + tilt away + gaze-away	8	s. shake + gaze-towards + m. shake	6
gaze-towards + m. shake + gaze-away	8	**gaze-towards + retract back + gaze-away**	6
gaze-towards + gaze-away + nod-down	8	s. nod + gaze-towards + gaze-away	5
gaze-away + m. nod + gaze-towards	8	**gaze-towards + tilt towards + gaze-away**	5
gaze-towards + gaze-away + s. nod	7	*gaze-away + s. nod + gaze-towards*	5
gaze-towards + nod-up + gaze-away	7	nod-up + gaze-towards + gaze-away	5

Indeed, in the most common trigrams, we see a common repeated pattern, that of the speaker gazing towards the listener, producing some kind of head gesture, then looking away, all in the span of a second (these are in bold). There are also instances of the opposite pattern, which is less common: gazing away, producing a head gesture, and then gazing towards (these are in italics). The most frequent head gestures here are nods. If this pattern is primarily being used by speakers to solicit a response and check to see if it is offered, then this suggests that this may be one function of nods (possibly a function that is dependent on this particular multimodal construction).

Listener bigrams are shown in Table 29.

Table 29. Listener Gaze and Head bigrams (1-second window)

Bigram	Frequency	Symmetric CP	CP: 2|1	CP: 1|2
nod-down + gaze-away	24	0.021	0.055	0.381
nod-up + gaze-towards	23	0.015	0.050	0.299
s. nod + gaze-away	33	0.012	0.075	0.161
m. nod + gaze-away	36	0.006	0.082	0.070
gaze-away + nod-down	12	0.005	0.190	0.027
gaze-towards + m. nod	33	0.005	0.064	0.072
gaze-away + m. shake	7	0.004	0.250	0.016
s. jut + gaze-towards	3	0.003	0.007	0.500
gaze-away + tilt away	7	0.003	0.175	0.016
s. nod + gaze-towards	16	0.003	0.035	0.078
tilt away + gaze-towards	7	0.003	0.015	0.175
gaze-away + s. nod	15	0.003	0.073	0.034

For listeners, we also see that nods tend to occur in proximity with gaze-shift. Here, however, the two most strongly associated bigrams are not necessarily components of larger trigram patterns, but are the result of a tendency for listeners to nod-down and gaze away at the same time, and nod-up and gaze-towards together as well. However, we will still look at the table of trigrams to compare listeners with speakers (Table 30).

Table 30. Most frequent Listener Gaze and Head trigrams (1-second window)

Trigrams	Freq.	Trigrams	Freq.
s. nod + gaze-away + gaze-towards	8	m. nod + gaze-away + gaze-towards	3
gaze-away + s. nod + gaze-towards	7	m. nod + gaze-towards + gaze-away	3
gaze-towards + nod-down + gaze-away	5	**gaze-towards + s. nod + gaze-away**	3
gaze-towards + m. nod + gaze-away	4	*gaze-away + m. nod + gaze-towards*	3
gaze-away + nod-up + gaze-towards	3	s. nod + s. nod + gaze-away	3

We see both patterns in listeners, more or less equally likely. The pattern of looking towards, producing a gesture, and returning is used (bold), but equally likely is the pattern of looking away, producing a gesture, and returning, but the low token frequencies don't really allow us to speak of patterns here.

4.3 Window Histograms

The window histograms for head and gaze are not included, as they either show normal or uniform distributions, or have too few tokens to discuss a clear pattern.

5. Speech+Gaze

5.1 Likelihood Measures

We saw in section 4.4 that speakers are more likely to produce head gestures while looking at listeners, possibly to look for a response. Yet we know that speakers spend most of their time speaking, and much less of their time looking at the listener. It they are looking for a visible response, will this motivate more co-gaze speech for some kinds of speech than others, and will this be as much of a consideration for listeners? We will look at likelihood measures of the overlaps between gaze to explore these possibilities, looking first at the overlaps between speaker gaze and speaker speech types (Table 31).

Table 31. Conditional Probabilities and odds ratios of Speaker gaze-towards with Speaker speech-types

	Cond. Prob. / Gaze-towards	**Cond. Prob. / Speech**	**Gaze-towards Odds-ratio**
Back-channel	0.016	0.546	**2.597*****
Interrogative	0.129	0.387	**1.398***
Declarative	0.542	0.337	**1.183***
Incomplete	0.056	0.196	0.493***
Filler	0.007	0.084	0.190***

We see that declaratives and interrogatives overlap with two thirds of speaker gaze-towards (0.542 + 0.129), while back-channels account for less than 2 percent. In the other direction, gaze overlaps with over half of speaker back-channels, but only around a third of speaker interrogatives and declaratives (more for interrogatives). Indeed, looking at the odds ratios, we see that speakers are much more likely than expected to look at the listener while the speaker is back-channeling, and only a bit more likely during the speaker's declaratives and interrogatives. They are much less likely to gaze-towards during incomplete and filler speech segments. As might be expected, it does seem to be the case that, in speech turn types, the strongest association is between gaze-towards and interrogatives, which is the most 'solicitatory' speech type.

Table 32 shows the overlaps of speaker gaze and speaker back-channels. There are only sufficient tokens for affirmations and laughs.

Table 32. Conditional Probabilities and odds ratios of Speaker gaze-towards with Speaker back-channels

	Cond. Prob. / Gaze-towards	Cond. Prob. / Speech	Gaze-towards Odds-ratio
Affirmation	0.011	0.550	**2.628*****
Laugh	0.004	0.400	**1.424***

We see that both kinds of speaker back-channels are more likely than expected to occur during gaze-towards, although only around half of all affirmations are co-gaze, and only 40% of all laughs.

We now shift to look at listeners. Table 33 shows likelihood measures for listener gaze-towards overlaps with speech types.

Table 33. Conditional Probabilities and odds ratios of Listener gaze-towards with Listener speech-types

	Cond. Prob. / Gaze-towards	Cond. Prob. / Speech	Gaze-towards Odds-ratio
Back-channel	0.055	0.768	0.564*
Interrogative	0.017	0.700	0.403**
Declarative	0.024	0.576	0.222***
Incomplete	0.002	0.447	0.142***

Looking at conditional probabilities, we see that listener speech-types also mainly overlap with gaze-towards – proportions of all speech types are much higher than for speakers. Still, all behaviors are less likely to overlap than expected, with back-channels at least showing greatest likelihood, followed by interrogatives. Listeners tend to view turn-taking as an opportunity to glance away, while speakers view back-channels as a chance to glance towards.

Table 34 breaks listener speech into subtypes of back-channel.

Table 34. Conditional Probabilities and odds ratios of Listener gaze-towards with Listener back-channels

	Cond. Prob. / Gaze-towards	Cond. Prob. / Speech	Gaze-towards Odds-ratio
Continuer	0.007	0.921	**2.073****
Acknowledgment	0.016	0.844	ns.
Collaborative finish	0.003	0.782	0.636*
Affirmation	0.004	0.769	0.590**
Assessment	0.026	0.709	0.418***
Laugh	0.012	0.697	0.400**
Newsmarker	0.002	0.617	0.285***

The bulk of listener back-channels overlap with gaze-towards, but this is especially so for continuers. They are the only behavior more likely than expected to occur during gaze-towards, although acknowledgments are at chance, so they at least do not seem to be repelled by gaze. Others are less likely than expected, especially assessments, laughs, and newsmarkers. Each of these (along with affirmations) indicates a certain kind of processing of speech content. Affirmations respond to a polar question; assessments give an affective personal response to content; laughs indicate recognition of a subversion of expectations; newsmarkers indicate assimilation of new information. Continuers, of all the back-channel types, have little to do with comprehension, and more to do with signaling that the speaker should continue. One possibility is that looking away could also be a signal to the speaker that is interpreted as indicating listener engagement with the speaker.

5.2. N-grams

Bigram sequences of speaker gaze and speech types are presented in Table 35.

Table 35. Speaker Gaze and Speech type bigrams (1-second window)

Bigram	Frequency	Symmetric CP	CP: 2|1	CP: 1|2
gaze-away + declarative	216	0.045	0.205	0.222
declarative + gaze-towards	217	0.045	0.219	0.206
interrogative + gaze-towards	69	0.022	0.070	0.321
gaze-away + filler	61	0.015	0.232	0.063
gaze-away + incomplete	69	0.014	0.201	0.071
declarative + gaze-away	115	0.013	0.118	0.109
back channel + gaze-away	30	0.009	0.031	0.306
gaze-towards + declarative	99	0.009	0.094	0.100
incomplete + gaze-towards	53	0.008	0.054	0.154
filler + declarative	46	0.008	0.044	0.175
gaze-away + interrogative	38	0.007	0.177	0.039
gaze-towards + back channel	23	0.005	0.235	0.023

The strongest trend in these bigrams is for speech onsets to follow gaze-away and precede gaze-towards, much as we saw with speaker gaze and head gesture. However, following from the greater overlap between gaze-towards with back-channels and interrogatives (seen in Table 33), we see that back-channels and interrogatives are less likely to follow gaze-away than other speech types. In fact, it is more common to see back-channels follow gaze-towards, or gaze-towards to follow interrogatives. For back-channels, this is likely because back-channels are in response to listener speech (a time when speakers tend to gaze towards listeners), and for interrogatives, this is likely because speakers are soliciting a response, and want to be able to see it.

Table 36 shows bigram gaze and speech pairs for listeners.

Table 36. Listener Gaze and Back-channel onset bigrams (1-second window)

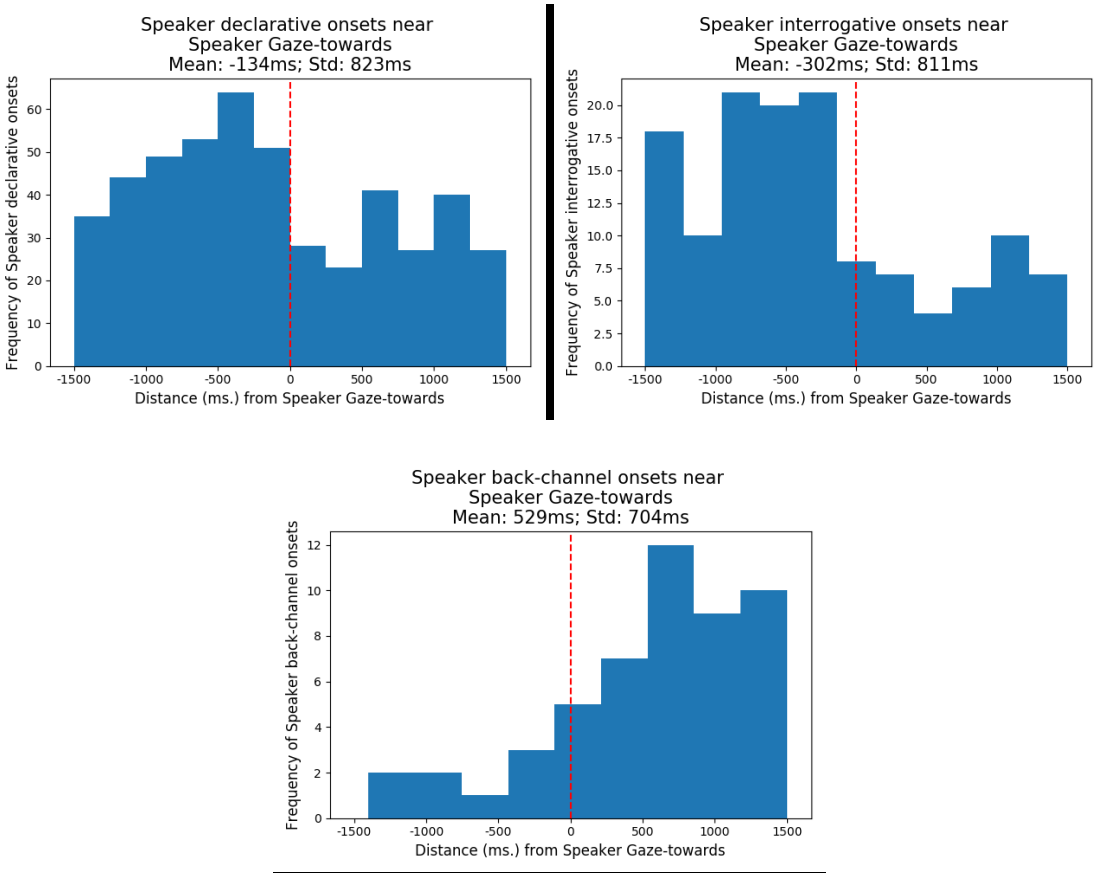
Bigram	Frequency	Symmetric CP	CP: 2|1	CP: 1|2
gaze-away + assessment	38	0.015	0.171	0.086
laugh + gaze-away	18	0.006	0.041	0.148
assessment + gaze-away	24	0.006	0.055	0.108
newsmarker + gaze-away	7	0.005	0.016	0.333
assessment + gaze-towards	22	0.005	0.048	0.099
gaze-towards + laugh	14	0.004	0.115	0.031

Unlike with speakers, almost all listener bigrams associate speech onsets with gaze-away, either before or after. Although they are not the most frequent back-channel, assessments are the most strongly associated with gaze-shift, behaving similarly to turns in this way. Acknowledgments, despite being the most frequent back-channel type, do not appear here at all, and nor do highly frequent continuers. These apparently are not commonly co-produced with gaze-shift.

5.3 Window Histograms

Only speaker turns and gaze-shifts show enough tokens to provide clear window histograms.

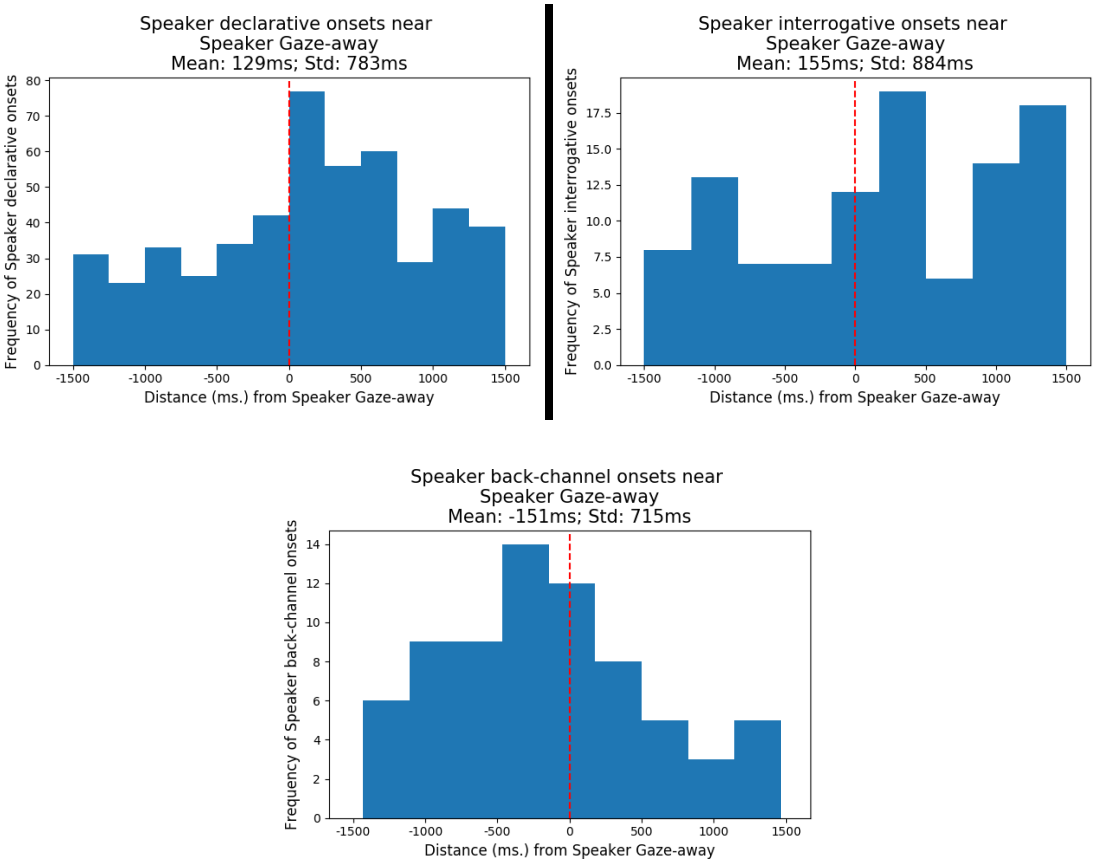
Figure 42. Window histogram – Speaker speech onsets near gaze-towards



The above figures show speaker declaratives, interrogatives, and back-channels occurring near speaker gaze-towards. We see a strong tendency for these to precede the gaze-shift for declaratives and interrogatives, and a tendency to follow it for back-channels. For back-channels, the gaze-towards tends to be timed more with the listener speech turn that preceded the speaker back-channel. It is less clear why speaker declaratives and interrogatives should start before the speaker gazes towards the listener, rather than after or concurrently. One (very tentative) possibility is that speakers gaze away from listeners' faces because they want to reduce cognitive processing load – faces carry a great deal of information that is unconsciously processed, and they have finite cognitive resources for doing speech production. So speakers would look away while planning their

speech, and once the speech production was planned and underway, they look back at listeners.

Figure 43. Window histograms – Speaker speech onsets near gaze-away



We see the opposite patterns for these three speech types relative to gaze-away. Declaratives and interrogatives are more likely to follow gaze-away, while back-channels are more likely to precede them, although the strength of the tendencies are not as pronounced as those relative to gaze-towards.

6. Summary and Hypotheses

6.1 Summary

In this chapter, we looked at the timing relations between behaviors of different modalities within an individual participant. Section 2 looked broadly at all four modalities, without distinguishing by speech or head subtypes. We saw different patterns of dependency across modalities for speakers and listeners. While speech, manual, and head gesture overlapped at greater than random chance for both roles, these three modalities were much less likely to co-occur during gaze-towards for listeners. For speakers, these three modalities were close to being at chance with gaze-towards, except for head gesture, which was more likely to co-occur with gaze-towards than expected. Looking at proximal behavior onsets, we saw close temporal timing of speaker and listener modalities, with speakers showing greater temporal coordination between gaze-towards and other modalities, and listeners showing more precise coordination between gaze-away and other modalities.

In Section 3, we looked in detail at the timing between head and speech subtypes. We saw strong association between most head gestures and speech types, but different categories of head gesture were attracted to and repelled from different speech types. Speaker interrogatives attracted ‘motion-towards’ head gestures and repelled ‘motion-away’ head gestures. Speaker fillers repelled all types of head gestures, but repelled half-cycle, repositioning gesture the least. Listener acknowledgments also attract forward-motion gestures and repel away-motion, and listener affirmations attract both nods and shakes of the head. Looking at timing relations between head and speech onsets, we saw

that speaker head gesture onsets tend to slightly follow speech onsets, while listener head gestures slightly precede speech.

Section 4 looked at the relationship between head gestures and gaze. A major difference across roles was that, while co-gaze head gestures are common across roles, many speakers' head gestures are more dependent on gaze than expected, while only the multiple nod was at greater than chance overlap for listeners. A frequent trigram in speaker production was the sequence of shifting gaze towards the listener, producing a head gesture (most commonly a nod), and gazing away again, as though to solicit a visible response from the listener. Listeners sometimes exhibited this same pattern, but were equally likely to do the opposite: gaze away, produce a head gesture, then return their gaze to the speaker.

Section 6 looked at the timing relationships between speech and gaze. Differences in the association between these modalities were seen across roles, with listeners being less likely than expected to produce co-gaze speech, and speakers, and speakers being more likely to produce co-gaze speech for back-channels, interrogatives, and declaratives. For speakers, back-channel and interrogative onsets were more likely to co-occur with gaze-towards, while other speaker speech onsets were more likely to co-occur with gaze-away. For listeners, all back-channels were more likely to co-occur near gaze-away, particularly assessments.

6.2 Hypotheses

These results open the possibility for a number of hypotheses, which could be further tested with qualitative or experimental analysis.

Hypotheses and further research questions:

1. Listener head onsets tend to precede listener speech onsets. Is this equally true, regardless of whether the speaker is gazing towards them or still speaking?
    - Listener head nods tend to respond to speaker gaze shift, while listener speech tends to respond to speaker pauses. But when listeners produce a head nod and a spoken back-channel together, the corresponding speaker cues are not necessarily also produced together.
    - We might hypothesize that, following Growth Point Theory (McNeill 1992), a single source sends signals to the speech and gesture production systems and times them to align in a similar way, regardless of whether or not both the head nod and spoken back-channel are easily detectable – this would predict little variance in timing regardless of whether the speaker was looking at the listener or speaking. We might also predict that the timing between these co-produced behaviors is influenced by the window of availability, by whether or not one or the other of the two behaviors was easily detectable – this would predict that the timing between the listener head nod and speech would vary depending on whether there was a window of availability opening or closing nearby.
2. Laughs and shakes can both be responses to discomforting things, but they are almost never co-produced. Are they responding to different kinds of discomforting things?
    - One hypothesis might be that they respond to the same kinds of things (but are not co-produced, for some reason, possibly because they convey a different

kind of response), while another might be that they respond to different categories of discomforting information – this could be tested using the Fisher's Exact Test to examine overlaps. For example, laughs might respond to things that are both discomforting and surprising, while shakes might respond to things that are discomforting and sad.

3. Speaker head gesture overlaps more than expected with gaze-towards (1.52), while speech overlap with gaze-towards is at chance (0.98), and manual gesture is only slightly more likely (1.22). One possibility that could be explored is that some speaker head gesture is intended to elicit a response, and so the speaker looks to check that the signal has been delivered.
    - Here one might make a strong hypothesis that some speaker head gesture is designed to elicit a visible response, and it is more likely to be produced while the speaker is looking at the listener because the speaker wants to see the visible response. To test this, one could look for an effect in the listener. We can count the proportion of listener head gestures that follow speaker head gesture *during* mutual gaze compared to those that follow speaker head gesture *outside* of mutual gaze. If these are more frequent during mutual gaze, this would support the hypothesis.
4. Speaker multiple nods near back-channels and single nods near fillers both show a strong tendency to precede their accompanying speech onset. This is not the case for nods down near interrogatives, which tend to follow the speech onset.
    - This might be a fruitful area for close qualitative analysis, but here are some possible explanations. Listener multiple nods also tend to precede back-

channel onsets, so speaker behavior might be obeying the same conventions in this case. When nods and fillers co-occur, both may be serving to facilitate the upcoming speech.

5. Acknowledgments are more likely than expected to overlap with both head shakes and head nods, which are head gestures that tend to be interpreted as having very different functions. Do these head gestures correlate with acknowledgments of different kinds of speech content (such as positive or negative affect)?
    - This could be tested simply by counting the proportion of times acknowledgments followed speaker content that was positive or negative in affect. Other possible categories that might differentiate these responses could be given vs. new information, surprising vs. unsurprising information, or even declarative vs. interrogative statements.

Continuers are the only listener back-channel more likely than expected to overlap with listener gaze-towards. These are also the back-channels that have the least to do with comprehension, and more to do with signaling that the speaker should continue. For the other back-channels, it could be that looking away is also a signal that speech is being processed, which would account for the greater overlap with listener gaze-away.

- One could hypothesize that listener continuers co-occur with listener gaze because the listener is attending more to the available cues in the speaker than their content. This could be tested by examining the responsiveness to cues in continuers compared with other back-channels (e.g. do they co-occur in response to a greater number of cues, or are they more closely timed with

certain cues). It could also be tested by manipulating whether or not the listener can see the speaker, and counting the frequency of continuers – if they are less frequent when the listener can't see the speaker, this would suggest they are produced more in response to visible speaker cues.

6. Speaker half-cycle head gestures are the most likely to co-occur with speaker fillers. These gestures often signify a shift in perspective. It's possible that different kinds of half-cycle gestures correspond to different kinds of perspective shifts, which could be examined qualitatively. If so, this would have interesting implications for planning.
    - If could be that different kinds of half-cycle gestures correspond to different kinds of narrative shifts, such as taking a different viewpoint, reconsidering a previous statement, expanding on a previous statement, or moving on to a new topic. Associations between these could be found by looking at the odds ratios of the overlaps of different half-cycle gestures and different categories of narrative shifts.
7. For speaker non-speech segments, which consists of all frames in which no (speaker) speech is occurring, only multiple nods are more likely than expected. This is probably both because it is a common back-channel behavior (and general acknowledgment of an interlocutor behavior), and because it is a thinking behavior, or a behavior one does when one doesn't quite know what else to do.

CHAPTER VI: ACROSS-ROLE / WITHIN-MODALITY

1. Introduction

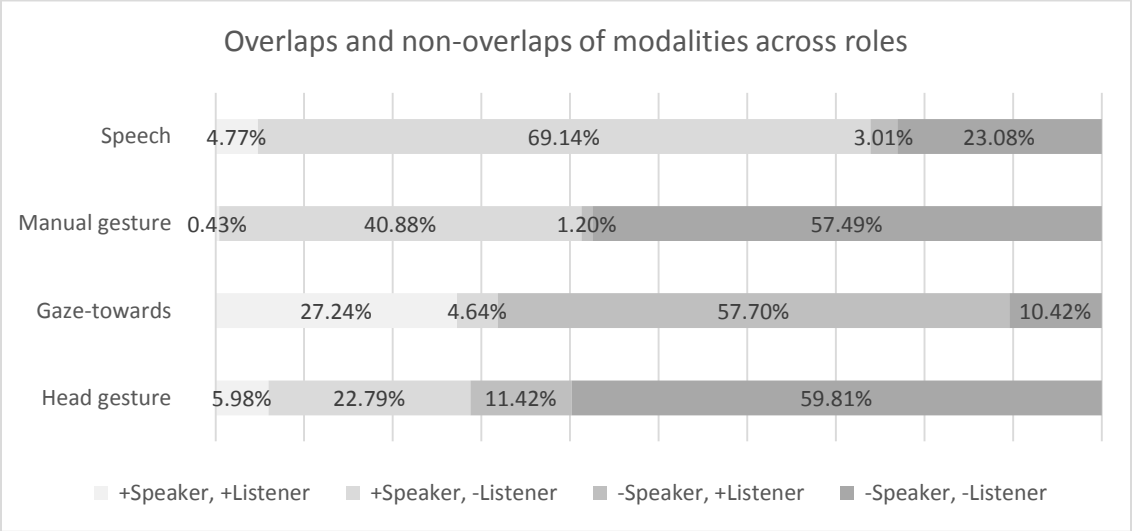
Chapter 6 looks at the timing relations of each of the four modalities across the roles of speakers and listeners. Section 2 will give an overview of the four modalities, and a more detailed analysis of the across role interactions between gaze and manual gesture, which are not coded for subtypes. Section 3 will examine the timing relations of subtypes of head gestures across roles, and Section 4 will examine the timing relations of subtypes of speech. Section 5 will summarize the findings in this chapter and lay out a set of hypotheses that could be tested in future qualitative analysis, based on these findings.

2. Within-mode Co-occurrence Patterns: Four Modalities

2.1 Likelihood Measures

In this section, we will look at the trends in overlapping behaviors of the same modality across roles, and the dependencies between them. To start out, we'll take a look at the nature of this overlap for each modality. Figure 44 shows the overlaps and non-overlaps each modality as proportions of the entire corpus. +Speaker, +Listener shows the percent of the corpus in which both speaker and listener were producing a behavior at the same time, +Speaker, - Listener shows the percent of the corpus in which speakers were producing a behavior and listeners weren't, and so on.

Figure 44. Overlaps and non-overlaps of modalities across roles



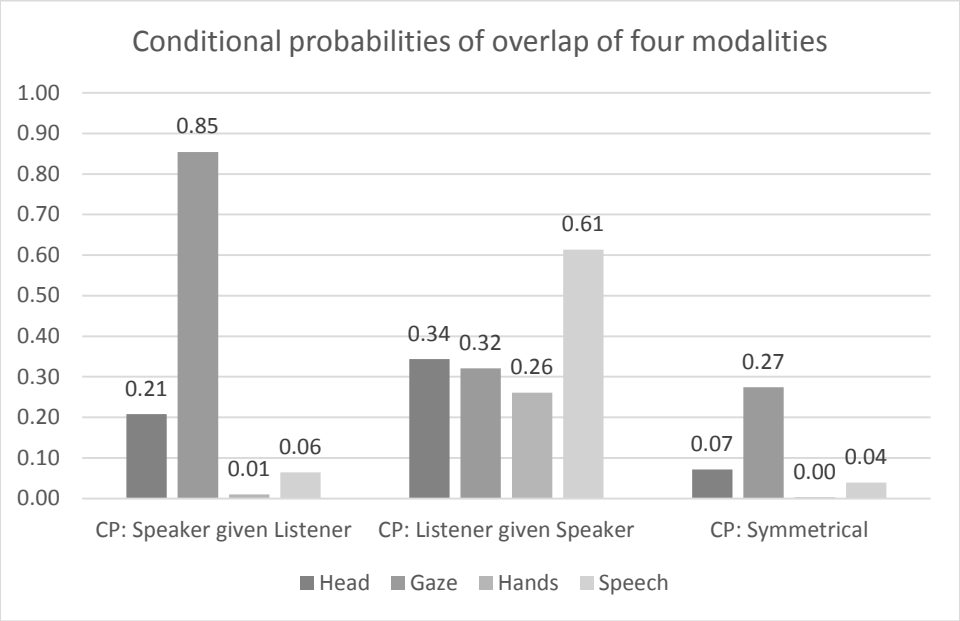
The first bar in each modality is the percent of mutual production. This is quite a small proportion of the time for most behaviors – very small for manual gesture (0.43%) and around 5 or 6% for speech and head gesture. Mutual gaze, as it is called in the literature (Goodwin 1980), is a larger percentage (27%), although it is less than half the percentage of time in which listeners are gazing-towards and speakers are not. Mutual gaze is the period of time when mutual negotiation can be done with non-verbal articulators, making it an important subset of this corpus. We will look more closely at this in Chapter 7, Section 2.

The second and third bars show the frames in which speakers are active and listeners are not, and when listeners are active and speakers are not. Unsurprisingly, for speech, manual, and head gestures, speakers' second bars are larger than their third bars, and the reverse is true for gaze-towards. The difference between these two bars is smallest for head gesture, where listeners are doing solo-head gesture at half the rate of speakers, much more similar than for speech (a rate of 3 to 69) and manual gesture (1.2 to 41). Of

all the modalities examined here, it could be argued that head gesture is where listeners are most active. The fourth bar is the time in which neither participant is active in a given modality – pauses between speech and gesture and periods of mutual gaze-away. Mutual gaze-away is the shortest bar in this category, followed by mutual speechlessness. Each of these bars, though, are made up of a great many small pauses and glances away – there are few long periods of silence and mutual gaze-away.

Having familiarized ourselves with these overlaps, we will take a look at the co-occurrence patterns of the four modalities with no categorization of subtypes. Figure 45 shows the conditional probabilities of each modality across roles. The first set of bars shows the number of overlapping frames of speaker and listener behaviors in each modality as a proportion of the total number of frames of the listener. The second set of bars shows the number of overlapping frames as a proportion of speaker frames. The third set of bars, the symmetrical conditional probability, is the product of the first two sets of bars, and gives a measure of the joint dependence of the behaviors on each other. For this first chart, to give a picture of the overall distributions, all within-modality behaviors are aggregated together, including non-congruent behaviors.

Figure 45. Conditional probabilities of overlap of four modalities



We see that, in terms of speaker modalities co-occurring with the same listener modalities, the vast majority of speaker gaze occurs while the listener is gazing towards them. Other modalities do not show the same degree of overlap. Twenty percent of speaker head gesture does overlap with listener head gesture, but there is relatively very little overlap of speaker hands, and, interestingly, only a small proportion of speaker speech overlaps with listener speech. We can compare this to the overlaps of modalities as a proportion of listener behavior, and see that there is more uniformity across modalities. Listener head gesture, gaze-towards, and manual gesture all overlap with these same speaker modalities to a similar degree. More than half of listener speech overlaps with speaker speech, however, predominantly back-channels, as we shall see shortly. Comparing across roles, we see the greatest similarity is the proportions of overlapping head gestures, further evidence of the temporal similarities between these

behaviors across roles, and the greatest disparities are in eye-gaze and speech, which are each very differently distributed across roles.

To get a sense of how much these overlaps are different from what we would expect from a chance distribution, we take a look at the odds ratios of each modality's observed overlap to the expected overlap, shown in Figure 46.

Figure 46. Odds ratio (log-transformed) of observed overlaps of four modalities (across role) to expected overlaps

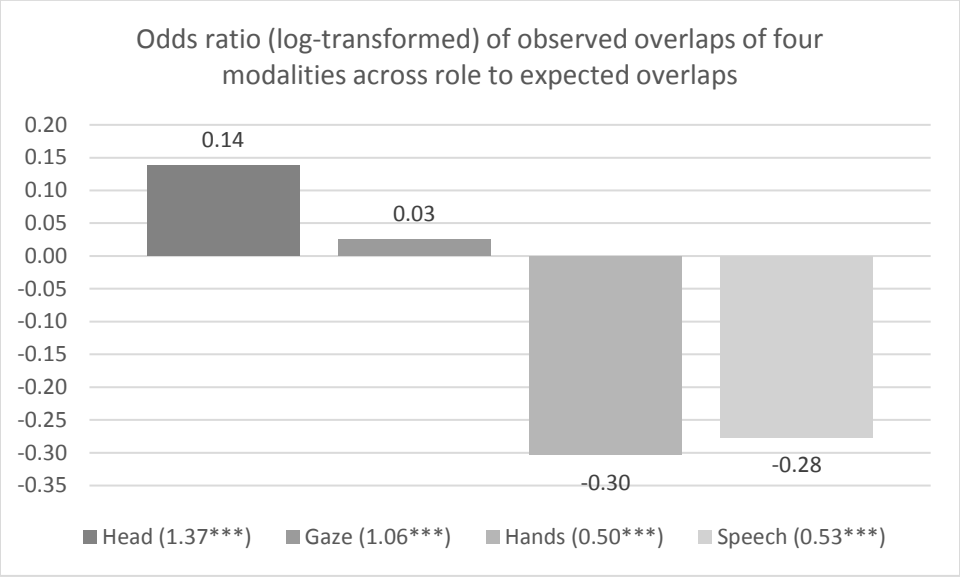


Figure 46 nicely illustrates the three kinds of results we can interpret from the odds ratio[24]. We see that the overlap of head behaviors across roles is greater than expected (137% of what we would expect from a chance distribution). While this may or may not be caused by one role or the other, it is definitely the case that speakers and listeners engage in simultaneous head gesture at a rate greater than expected from chance. For the overlap of gaze-towards across roles, we see that the overlap is 6% more than we would

[24] All reported odds ratios are significant at the level of $p < .001$, unless otherwise mentioned. Significance below .01 is marked with **, significance below .05 is marked with *, and anything else is marked with ns.

expect from chance. This is a significant difference ($p < .001$) from chance, but it is very close to an odds ratio of 1 (which is 0, when log-transformed), which expresses being at a chance distribution, and even though there is a significant difference, this difference is extremely small. This is especially interesting considering the complicated exchanges of gaze shifts that speakers and listeners engage in throughout the stories, and warrants closer examination later in this section. Finally, looking at the overlap of hands and speech across roles, we see that observed overlap of these modalities across roles is only half of the expected amount. For speech, this makes sense, as overlapping speech is much more likely to impede communication than overlapping heads. This same argument is less compelling for manual gesture, since gesture does not crowd the visual signal in the same way speech crowds the acoustic signal, but if we recall that all manual gesture in this corpus is co-speech gesture (and that there are relatively few listener manual gestures), then it makes sense that this odds ratio must follow from the odds ratio of speech overlaps.

We will now look at the overlaps of modalities across roles in closer detail. Since gaze and manual gesture are not coded for any subtypes, and since this chapter is limited to within-mode analyses, we will not be exploring these modalities in any detail, and will proceed with the co-occurrence patterns of subtypes of speech and head gesture.

2.2 N-grams (Gaze)

There are very few instances of listener manual gesture, and they do not distribute in any clear sequences with speaker manual gesture, so they will not be analyzed in this section. Gaze-shift across roles, on the other hand, is highly sequential. Table 37 shows the bigram frequencies of speaker and listener gaze-shifts, looking within 1-second

windows[25]. Since these categories are binary, conditional probability will not tell us anything more than frequency will, and so is omitted.

Table 37. Speaker and Listener Gaze bigrams (1-second window)

Bigram (1 + 2)	**Frequency**
L. gaze-away + **S. gaze-away**	86
S. gaze-away + L. gaze-away	78
L. gaze-towards + **S. gaze-towards**	76
S. gaze-towards + L. gaze-away	71
L. gaze-towards + **S. gaze-away**	68
S. gaze-away + L. gaze-towards	65
S. gaze-towards + L. gaze-towards	59
L. gaze-away + **S. gaze-towards**	57

Each permutation is relatively frequent even in this short window. Given how much less frequently listeners shift their gaze compared to speakers (4.62/min. for listeners; 10.13/min. for speakers) this suggests an interactive nature to these bigrams. The most frequent involve speakers and listener doing the same thing, either sequentially looking away or sequentially looking towards. The most frequent bigram involving different directions (S. gaze-towards + L. gaze-away) has speakers gazing at listeners, who then look away.

There are only eight possible types in this analysis, so it will be helpful to look at longer n-grams to see whether these bigrams are part of larger patterns. Table 38 shows the most frequent 3-grams (in a 2-second window) and 4-grams (in a 3-second window).

[25] These only include *across*-role bigrams, not *within*-role bigrams. For comparison, though, the within-role bigram frequencies are: S. gaze-towards + S. gaze-away (329), S. gaze-away + S. gaze-towards (113), L. gaze-away + L. gaze-towards (104), and L. gaze-towards + L. gaze-away (61).

Table 38. Speaker and Listener Gaze 3-grams (2-second window) and 4-grams (3-second window)

3-gram	Freq.	4-gram	Freq.
L. gaze-away + **S. gaze-away** + L. gaze-towards	60	**S. gaze-towards** + L. gaze-away + **S. gaze-away** + L. gaze-towards	41
S. gaze-towards + L. gaze-away + **S. gaze-away**	50	L. gaze-away + **S. gaze-towards** + L. gaze-towards + **S. gaze-away**	31
L. gaze-towards + **S. gaze-towards** + **S. gaze-away**	46	L. gaze-away + **S. gaze-away** + L. gaze-towards + **S. gaze-towards**	28
S. gaze-towards + L. gaze-towards + **S. gaze-away**	44	**S. gaze-towards** + S. gaze-away + L. gaze-away + L. gaze-towards	25
S. gaze-away + L. gaze-away + L. gaze-towards	41	**S. gaze-away** + L. gaze-away + **S. gaze-towards** + L. gaze-towards	21
L. gaze-away + **S. gaze-towards** + L. gaze-towards	41	L. gaze-away + L. gaze-towards + **S. gaze-towards** + **S. gaze-away**	21
S. gaze-towards + **S. gaze-away** + L. gaze-away	41	**S. gaze-away** + L. gaze-away + L. gaze-towards + **S. gaze-towards**	19

The most frequent 3-grams and 4-grams share a similar pattern of exchanging gaze, speakers gazing towards the listener, listeners gazing away, speakers gazing away again, and listeners gazing back at the speaker, or the same pattern with the roles reversed. A bit less frequent are sequences of within-role gaze-shift bigrams, speakers gazing towards then gazing away, followed by listeners gazing away, then gazing towards. The sequences allow for mutual gaze, but they don't seem designed to maintain it. The fact that gaze-away + gaze-away bigrams are more frequent in the 1-second window than gaze-towards + gaze-towards suggests that participants are more interested in getting out of the mutual gaze than getting into it (and this is more true for listeners than speakers: in gaze-away bigrams, listeners tend to gaze-away first, and in gaze-towards bigrams, speakers are more likely to gaze-towards second). However, the difference between 86 and 76 is slight enough that we should hesitate to draw conclusions from this.

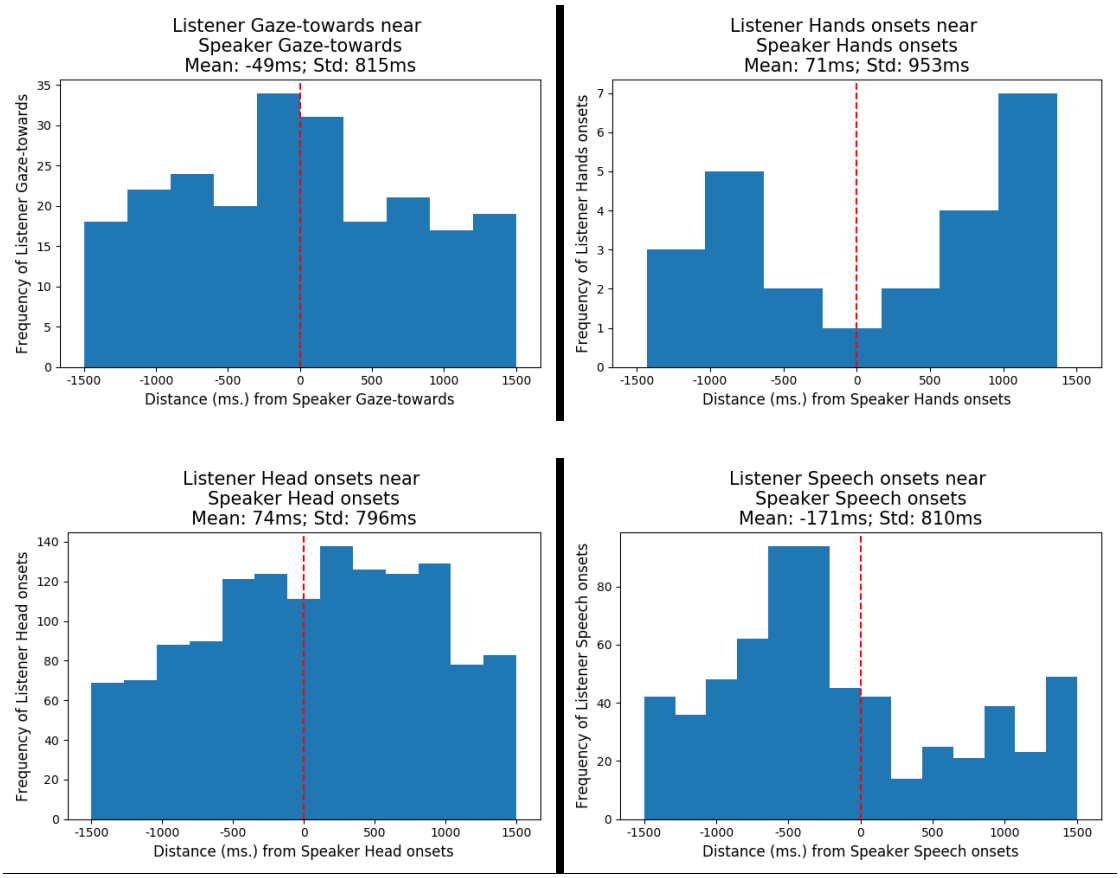
2.3 Window Histograms

We will look at histograms of heads and speech in more detail in Sections 3 and 4, but here we will take a look at the timing of the onsets and offsets of each broad modality

across roles. In Chapter 5, we only looked at onsets near onsets, but listeners and speakers are more likely to time the onsets of their behaviors with the offsets of their interlocutor's behaviors, so here we will look at interactions between onsets and offsets, as well. Manual gesture is included for completeness' sake, but there are too few proximal tokens to see any clear patterns.

2.3.1 Listener Onsets Near Speaker Onsets

Figure 47. Window histogram – Listener modality onsets near Speaker modality onsets

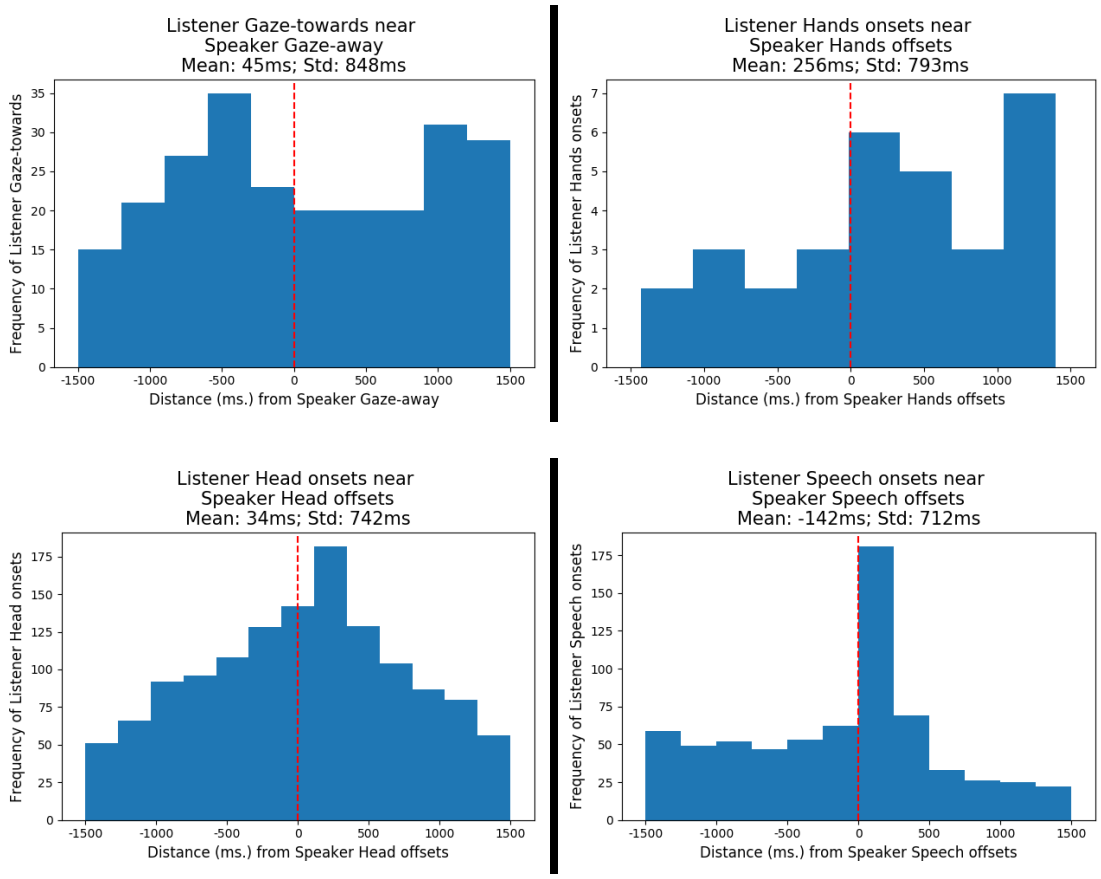


We begin by looking at onsets near onsets. Gaze-towards onsets have a slight peak near each other, but on the whole are mostly uniformly distributed within this window. Head onsets tend to gently peak near each other, but are also likely to occur on either side. Speech shows the clearest relationship, with listener speech onsets tending to occur

before speech onsets (but see listener speech offsets near speech onsets). Overall, onsets do not seem to be closely timed to occur with other onsets within the same role.

2.3.2 Listener Onsets Near Speaker Offsets

Figure 48. Window histogram – Listener modality onsets near speaker modality offsets

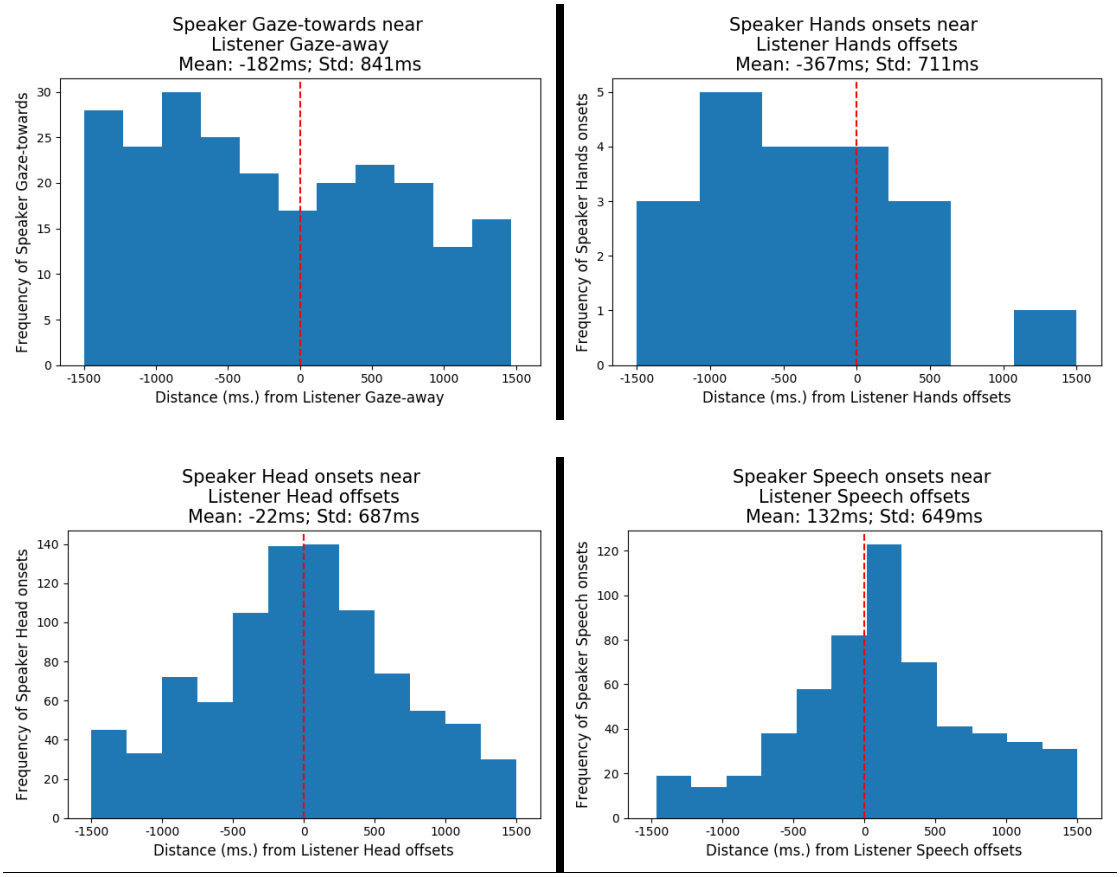


Listener gaze-towards onsets are mostly uniform, with a slight peak just before speaker gaze-away. Listener manual gesture onsets *may* be more likely to occur after speaker manual gesture offsets. Listener head onsets are nearly symmetrically likely to occur near speaker head offsets. There is an extremely sharp peak in listener speech onsets immediately following speaker speech offsets, showing that listeners are very good at identifying the end of a speech segment (relying, no doubt, on prosodic, syntactic, and pragmatic features in the speech).

Here we see a greater tendency for listener onsets to be timed relative to speaker offsets of the same role. This is most clearly evident in the speech modality (and, by extension, in manual gesture), but is also evident in head gesture.

2.3.3 Speaker Onsets Near Listener Offsets

Figure 49. Window histograms – Speaker modality onsets near Listener modality offsets



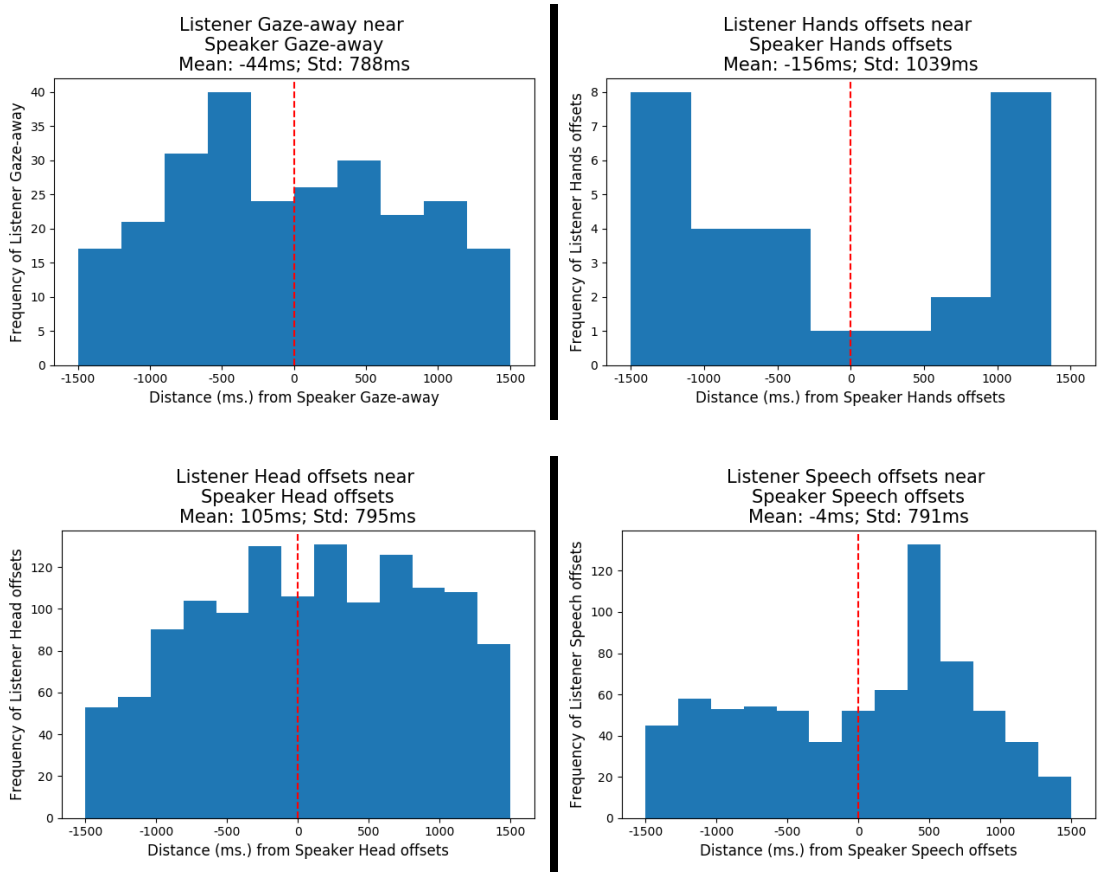
Speaker gaze-towards onsets are more likely to occur before listener gaze-away, as we saw in Section 2.2. Speaker head onsets show similar symmetrical rise in likelihood from both directions, much like we saw in Sections 2.3.1 and 2.3.2 – this symmetry across roles suggests that neither speakers nor listeners are the sole leader in these interactions. Speaker speech onsets also show the same peak as listener speech onsets, but with more

gradual rising and tapering before and after, suggesting that speakers are less interested in accommodating listener speech than vice versa.

For speaker onsets near listener offsets, we see some of the same patterns as in Section 2.3.2, in head gesture and speech. There is a lower token count for speaker head gesture onsets, but the frequency distributions is quite similar. For speech, the overall pattern is the same as for listener speech onsets, but speaker onsets do not show the same precision in timing.

2.3.4 Listener Offsets Near Speaker Offsets

Figure 50. Window histograms – Listener modality offsets near Speaker modality offsets



Listener gaze-away tends to precede speaker gaze-away, also seen in Section 2.2. We do not see much of a pattern with head gesture offsets, only a slight rise in likelihood approaching the intersection. For listener speech offsets, however, there is a peak around 500ms after the offset of speaker speech, which corresponds roughly to the average length of a back-channel.

As with onsets and onsets, offsets do not seem to be timed relative to each other within modality, and cases where patterns exist (such as with speech offsets) actually follow from onset and offset timing relations.

3. Speaker Head Subtypes + Listener Head Subtypes

3.1 Likelihood Measures

3.1.1 Conditional Probabilities

We turn now to the co-occurrence patterns across roles in the head modality. We saw in Figure 46 that head behaviors were the only modality where there was substantially greater overlap across roles. Although, as with speech, when we look more closely at specific subtypes of this modality, we see variation in which kinds of behaviors tend to co-occur with other behaviors.

Many head subtypes are infrequent enough, compared to the frequency of nods and shakes, that showing conditional probabilities of all behaviors will not be very informative. Table 39 shows the conditional probabilities of the overlaps of listener head behaviors with speaker multiple nods, single nods, and multiple shakes (as proportions of listener behaviors), which account for the majority of overlap with listener behaviors.

Table 39. Conditional probabilities of speaker multiple nods, single nods, and multiple shakes, given listener head behaviors

Speaker	Listener head behavior						
	M. nod	**S. nod**	**Nod up**	**Nod down**	**S. shake**	**M. shake**	**Tilt t.**
M. nod	0.133	0.060	0.058	0.109	0.174	0.067	0.143
S. nod	0.052	0.059	0.076	0.031	0.013	0.052	0.062
M. shake	0.082	0.060	0.024	0.011	0.051	0.078	0.112
	Tilt a	**Tilt t+r**	**Tilt a+r**	**Jut in**	**S. jut**	**Retr back**	**S. retr**
M. nod	0.153	0.192	0.133	0.009	0.088	0.062	0.000
S. nod	0.044	0.026	0.041	0.090	0.056	0.013	0.026
M. shake	0.044	0.048	0.000	0.110	0.000	0.086	0.086

Speaker multiple nods have the greatest overlap with speaker head gestures, being particularly high for listener multiple nods (0.13), single shakes (0.17), and all four

subtypes of tilts (between 0.13 and 0.19). Speaker single nods account for less than half the overlap of multiple nods overall, but for more overlap of juts-in (0.09). Speaker multiple shakes also account for less overlap than multiple nods, but for considerable overlap with tilts-towards (0.11), juts-in (0.11), retractions-back (0.09), and single retractions (0.9), all head gestures involving motion on the axis between the two participants. Speaker shakes often co-occur with speech describing unpleasant or negative events, and these towards and away-motion listener head gestures can effectively express appropriate responses of shock, amazement, or disbelief.

Listener multiple nods make up the majority of listener head gestures, and overlaps with other listener head gestures are not frequent enough to account for informative proportions of speaker head gestures. Table 40 shows the overlaps of listener multiple nods with speaker head behaviors as a proportion of speaker head behaviors.

Table 40. Conditional probabilities of listener multiple nods, given speaker head behaviors

Listener	**Speaker head behavior**								
	M. nod	**S. nod**	**Nod up**	**Nod down**	**S. shake**	**M. shake**	**Tilt t**	**Tilt a**	**Tilt t+r**
M. nod	0.214	0.121	0.109	0.118	0.110	0.140	0.106	0.105	0.095
	Tilt a+r	**S. wag**	**M. wag**	**Jut in**	**S. jut**	**M. jut**	**Retr back**	**S. retr**	
M. nod	0.109	0.094	0.182	0.085	0.210	0.288	0.089	0.153	

Listener multiple nods account for a substantial portion of overlap with most speaker head behaviors, but especially with multi-cycle behaviors, such as speaker multiple nods (0.21), multiple shakes (0.14), multiple wags (0.18), and multiple juts (0.29). Others have looked at the rhythmic, interactive nature of cycles of head nods, looking in detail at the

peaks of prominence, and finding a tendency for conversational participants to synchronize these head gestures (Louwerse et al, 2012).

3.1.2 Odds-ratios

Table 41 shows the odds ratios of observed to expected overlaps for subtypes of speaker nods, shakes, and wags with all listener head gesture subtypes (except wags, multiple juts, and multiple retractions, for which there were fewer than 10 tokens). Given the number of overlapping categories, this analysis will look first at speaker categories, then at listener categories.

Table 41. Odds ratios of observed overlaps (all listener head behaviors with speaker nods, shakes, and wags) to expected overlaps

Listener	**Speaker: nods, shakes, and wags**							
	M. nod	**S. nod**	**Nod-up**	**Nod-down**	**M. shake**	**S. shake**	**S. wag**	**M. wag**
M. nod	**2.46*****	**1.15*****	ns	**1.12***	**1.39*****	ns	ns	**1.86*****
S. nod	ns	**1.31*****	ns	0.60***	ns	**1.36****	0.37**	0.35***
Nod-up	ns	**1.73*****	**1.69****	ns	0.37***	0.05***	0.00**	0.00***
Nod-down	**1.73*****	0.67*	**1.79*****	ns	0.16***	0.48*	0.00**	**2.07****
M. shake	ns	ns	ns	ns	**1.27***	ns	0.00**	0.00***
S. shake	**2.98*****	0.28***	ns	0.00***	ns	0.23***	ns	0.00*
Tilt towards	**2.37*****	**1.38***	3.36***	**2.13*****	**1.89*****	ns	0.00*	ns
Tilt away	**2.55*****	ns	ns	0.00***	0.68*	ns	ns	**1.81***
Tilt t+r	**3.36*****	0.56***	**1.84***	ns	ns	**2.56*****	ns	0.00*
Tilt a+r	**2.17*****	ns	0.00***	ns	0.00***	0.00***	ns	**7.37*****
Jut in	0.12***	**2.06*****	0.00**	0.00***	**1.85*****	0.00**	ns	ns
S. jut	ns	ns	0.00***	0.00***	0.00***	0.00***	ns	ns
S. retr	0.00***	ns	0.00**	**1.99****	ns	**3.36*****	ns	ns
Retr. back	ns	0.27***	ns	**2.50*****	**1.40***	ns	**3.54*****	0.00***

Looking first at speaker nods, we see that multiple nods are more likely than expected to overlap with half of all listener head subtypes, and only less likely to overlap with two of them. As was suggested by Table 39, all forms of listener tilts are more likely than

expected to overlap (between 2.37 and 3.36 more likely) with speaker multiple nods. If this were our only set of data points, one might conclude that these different forms of tilts are not functionally different from each other, but we will see that this is not the case when looking at their overlaps with other behaviors. We also see that these multiple nods are more than twice as likely to overlap with listener multiple nods (2.46).] They are less likely than expected to overlap with juts-in (0.12) and single retractions (0.00), and they are at chance with single juts and retractions-back, all the head behaviors involving linear motion on the z-axis.

Speaker single nods are at least somewhat more likely than expected to overlap with listener single nods, as well as all other forms of listener nods except nods-down (0.67). They differ from multiple nods in certain ways: they are more likely to overlap with juts-in (2.06) and much less likely to overlap with retractions-back (0.27); they are also much less likely to overlap with single shakes (0.28), and they do not show the same degree of overlap with tilts.

Speaker nods-up show some interesting overlap patterns as well. They also are more likely than expected to overlap with listener nods-up (1.69). They are also more likely than expected to overlap with nods-down (1.79), as well as tilts-towards, both half-cycles (3.36) and full cycles (1.84), while they are much less likely to overlap with full cycles of tilts-away (0.00). Impressionistically, this often looks like speakers lifting their heads to start on a new perspective or portion of the story, and listeners tilting their heads towards to indicate either acknowledgment or a similarly shifted perspective. Speaker nods-up are much less likely to overlap with most listener linear z-axis head behaviors (juts and retractions).

Speaker nods-down, which often co-occur with the delivery of new and important or 'serious' information, seemingly as a way of emphasizing the content, are slightly more likely than expected to co-occur with listener multiple nods (1.15), but less likely to co-occur with single nods (0.60). Of all the nods, they are the only subtype not significantly more likely to co-occur with its counterpart from the interlocutor. They are also quite likely to overlap with tilts-towards (2.13, often interpreted as an acknowledgment), but very unlikely to co-occur with tilts-away (0.00). However, they are extremely unlikely to co-occur with either kind of jut (0.00), but quite likely to co-occur with both single retractions (1.99) and retractions-back (2.50). Despite the fact that retractions and tilts-away both involve movement away from the interlocutor, they tend to co-occur with distinct kinds of behaviors, with retractions co-occurring with surprise or amazed acknowledgment and tilts-away co-occurring with some kind of shifted perspective, and these co-occurrence patterns provide some formal evidence for this functional distinction.

The functions of speaker shakes, both single and multiple, are among the most difficult to define. They can co-occur with the telling of events that are negative in some way, but they can also occur in more neutral contexts, where they seem almost to be a response to the processing of something internal and possibly unsaid (almost like they are rejecting or dissociating from some aspect of the story). We will see in the second section of this chapter that the onset patterns of single and multiple shakes are quite similar, but their overlaps differ in some noticeable ways.

First, speaker multiple shakes are more likely than expected to overlap with listener multiple shakes (1.39), but speaker single shakes show the opposite pattern with listener single shakes (0.23). While both kinds of shakes are less likely than expected to overlap

with nods-up and nods-down, multiple shakes are more likely to overlap with multiple nods (1.39), while single shakes are more likely to overlap with single nods (1.36). Patterns are just as mixed in overlaps with tilts, where multiple shakes show a preference for tilts-toward (1.89, suggestive of an acknowledging response being elicited and/or given), and a dispreference for tilts-away (0.68), while single shakes are at chance for both. And while both are less likely to occur with full cycles of tilts-away, single shakes are much more likely to overlap with full cycles of tilts-towards (2.56). Finally, multiple shakes are more likely than expected to overlap with both juts-in (1.85) and retractions-back (1.40), while single shakes are only more likely than expected to overlap with single retractions (3.36).

Table 4.2 shows the odds ratios of speaker tilts, juts, and retractions with all listener head gestures.

Table 42. Odds ratios of observed overlaps (all listener head behaviors with speaker tilts, juts, and retractions) to expected overlaps

Listener	**Speaker: tilts, juts, and retractions**								
	Tilt towards	**Tilt away**	**Tilt t+r**	**Tilt a+r**	**Jut in**	**S. jut**	**M. jut**	**Retr. back**	**S. retr.**
M. nod	ns	ns	ns	ns	0.77***	**2.25*****	**3.39*****	0.81*	**1.51*****
S. nod	**1.61*****	ns	0.31***	0.56*	ns	0.70*	0.00***	**2.71*****	0.38***
Nod up	**3.47*****	0.36**	0.00***	ns	0.21***	0.19***	0.00*	0.37*	0.00***
Nod down	ns	ns	ns	ns	0.00***	0.12***	**5.22*****	ns	ns
M. shake	**2.02****	0.45***	0.00**	ns	**1.52***	0.00***	0.00**	0.62***	0.00***
S. shake	0.00***	ns	0.00	ns	0.00**	0.00*	ns	ns	**2.11***
Tilt t	ns	ns	ns	ns	ns	ns	**15.00*****	**3.04*****	0.00***
Tilt a	**2.29****	0.22***	0.00***	ns	0.00**	0.00***	ns	**2.07****	**4.18*****
Tilt t+r	0.00*	0.00***	ns	ns	0.00**	0.00*	ns	**3.68*****	0.00*
Tilt a+r	ns	ns	0.00**	**5.90*****	**2.56****	0.00***	ns	**2.68*****	**10.43*****
Jut in	0.00**	ns	**4.53*****	ns	0.00**	0.00*	ns	**4.37*****	ns
S. jut	0.00***	0.00**	**3.08*****	ns	0.00**	0.00**	ns	0.00*	0.00*
S. retr	**2.66**	0.00*	ns	ns	0.00***	ns	ns	**7.68***	ns
Retr back	ns	ns	ns	ns	0.00*	0.30*	ns	**2.11*****	ns

Between speaker and listener tilts, there is very little overlap. Tilts-away + return are substantially more likely to overlap than expected – listeners following after speakers, a case of imitation. The other case is when speakers tilt their heads towards the listener, and the listeners tilt their heads away (2.29). This interaction only occurs four times, but each time it is the listener that tilts away after the speaker has tilted towards, a pattern somewhat iconic of the direction of information sharing and receiving in a speaker-listener relationship.

3.2 N-grams

We look now at the sequential nature of head behaviors across roles. Because onsets and offsets are both potentially relevant boundaries in across-role timing relations, Table 43

includes both. Table 43 looks broadly at the onsets and offsets from each role that occur near each other.

Table 43. Speaker and Listener bigrams (1-second window)

Bigram (1 + 2)	Frequency	Symm. CP
S. Head offsets + L. Head onsets	324	0.049
L. Head offsets + **S. Head onsets**	267	0.033
S. Head offsets + L. Head offsets	234	0.025
L. Head offsets + **S. Head offsets**	226	0.023
L. Head onsets + **S. Head offsets**	218	0.022
L. Head onsets + **S. Head onsets**	214	0.021
S. Head onsets + L. Head onsets	213	0.021
S. Head onsets + L. Head offsets	208	0.020

Looking at which role is more likely to follow the other, we see what looks like greater responsiveness from listeners for most kinds of bigram pairs. However, the type of boundary-pair also seems to play a part in the degree of dependency. The offset+onset pair is most frequent, followed by offset+offset, regardless of role. Less dependent are the onset+offset and onset+onset pairs, suggesting that, if speakers and listeners are timing their heads to occur near each other's boundaries, they are timing their onsets relative to the other's offsets more than the other's onsets.

Because an individual head gesture consists of an onset and an offset, if we want to see the full interaction between a speaker and a listener head gesture, we need to also look at 4-grams. In Table 44 (4-grams), only 4-grams with a frequency of 6 or greater were included, and the window size was set at three seconds to match the one-second window size of bigrams (4-grams consisting of three consecutive bigrams).

Table 44. Listener and Speaker Head 4-grams (3-second window)

Type	4-gram	Freq.
S + L (overlap)	**S. m. nod onset +** L. m. nod onset + **S. m. nod offset** + L. m. nod offset	19
S + L (separate)	**S. s. nod onset + S. s. nod offset** + L. m. nod onset + L. m. nod offset	12
S + L (separate)	**S. nod d onset + S. nod d offset** + L. m. nod onset + L. m. nod offset	11
L + **S** (separate)	L. m. nod onset + L. m. nod offset + **S. nod d onset + S. nod d offset**	8
L + **S** (separate)	L. s. nod onset + L. s. nod offset + **S. s. nod onset + S. s. nod offset**	8
S + L (separate)	**S. tilt t onset + S. tilt t offset** + L. m. nod onset + L. m. nod offset	7
S + L (separate)	**S. nod d onset + S. nod d offset** + L. s. nod onset + L. s. nod offset	7
L + **S** (overlap)	L. m. nod onset + **S. s. nod onset** + L. m. nod offset + **S. s. nod offset**	6
S + L (overlap)	**S. s. nod onset** + L. m. nod onset + **S. s. nod offset** + L. m. nod offset	6
L + **S** (overlap)	L. m. nod onset + **S. m. shake onset** + L. m. nod offset + **S. m. shake offset**	6
L + **S** (separate)	L. m. nod onset + L. m. nod offset + **S. s. nod onset + S. s. nod offset**	6
L + **S** (separate)	L. s. nod onset + L. s. nod offset + **S. nod d onset + S. nod d offset**	6
S + L (separate)	**S. s. nod onset + S. s. nod offset** + L. s. nod onset + L. s. nod offset	6
S + L (separate)	**S. m. shake onset + S. m. shake offset** + L. m. nod onset + L. m. nod offset	6
L + **S** (overlap)	L. m. nod onset + **S. nod u onset** + L. m. nod offset + **S. nod u offset**	6
L + **S** (enclose)	L. m. nod onset + **S. nod d onset + S. nod d offset** + L. m. nod offset	6

The first thing to note is that these frequencies aren't especially large, considering the large numbers of listener and speaker heads. There are too many individual pairings to head gestures to look at in detail. However, all of these 4-grams involve two completed head gestures (the fact that none of these begin with an offset suggests that these pairs of head gestures are not merely random onsets and offsets, but coordinated pairs of head gestures). And nearly all are distributed into one of four categories, defined by whether the initial head onset is produced by the speaker or listener, and whether the two head gestures are overlapping or separate. As we can see, exactly half of these 4-gram types begin with speaker head gesture, and half with listener gesture (although speaker-initiated head gestures are more frequent, token-wise). Ten of the 4-grams involve two separate gestures, one after the other, and five involve overlap (all overlaps involve listener

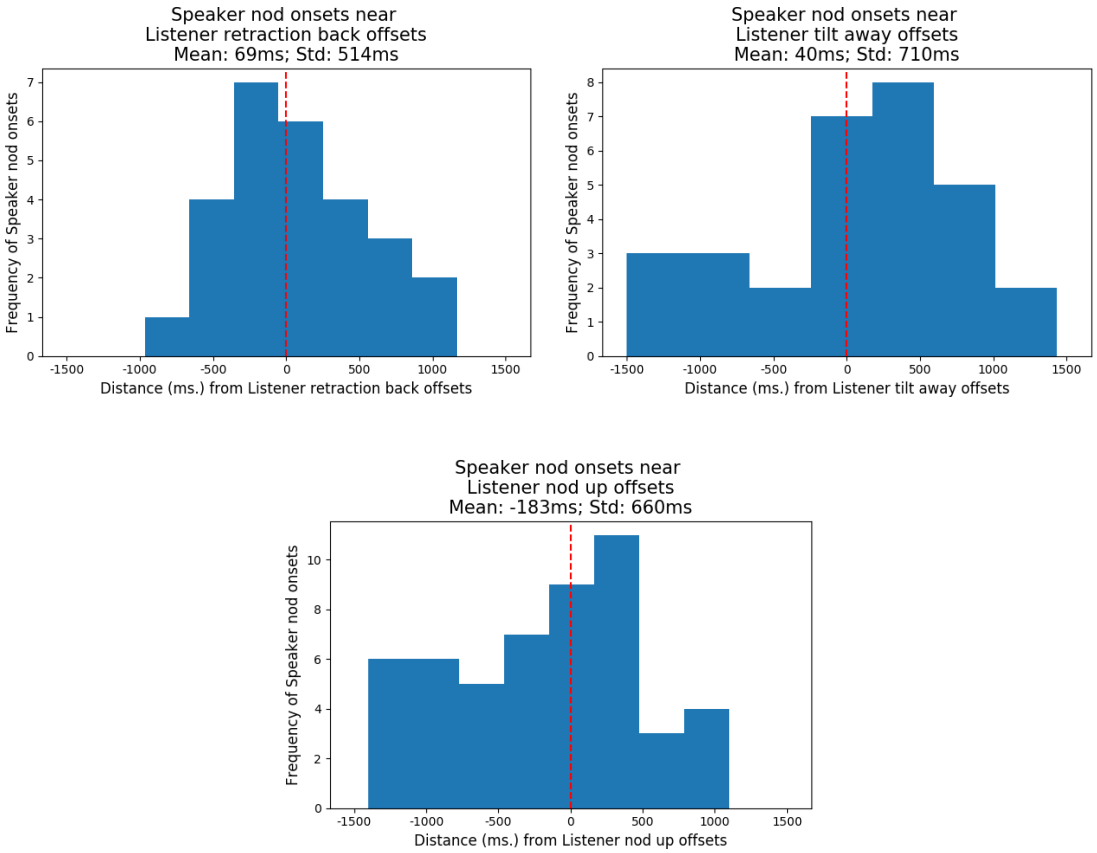
multiple head nods). Finally, one type of 4-gram had listeners beginning a multiple nod and speakers starting and finishing a single nod before the listener finished.

While speakers do tend to initiate these interactive pairs of gestures, it is striking how often listeners are also leading. Adding this to the symmetry of the timing between head gesture onsets and offsets across roles seen in Section 2.3, this leads us to make a suggestion, which will be reiterated in the hypotheses in Section 5.2: it make be that head gesture is a channel in which speakers and listeners engage in a separate sort of turn-taking, partially dependent and partially independent of the speech channel. It certainly seems that both roles are quite responsive, and there is clearly a back-and-forth exchange of gestures.

3.3 Window Histograms

We turn now to window histograms of speaker and listener head gestures. We've seen in Section 2.3 that, while some head onsets are timed to occur near other head onsets, there is clearer timing of onsets relative to offsets, for both speakers and listeners, and that is what we will see in the following figures. Note, however, that many head subtypes are relatively infrequent (particularly for listeners) and only those with sufficient tokens will be examined. We'll look first at speaker head onsets near listener head offsets, collapsing all speaker heads together.

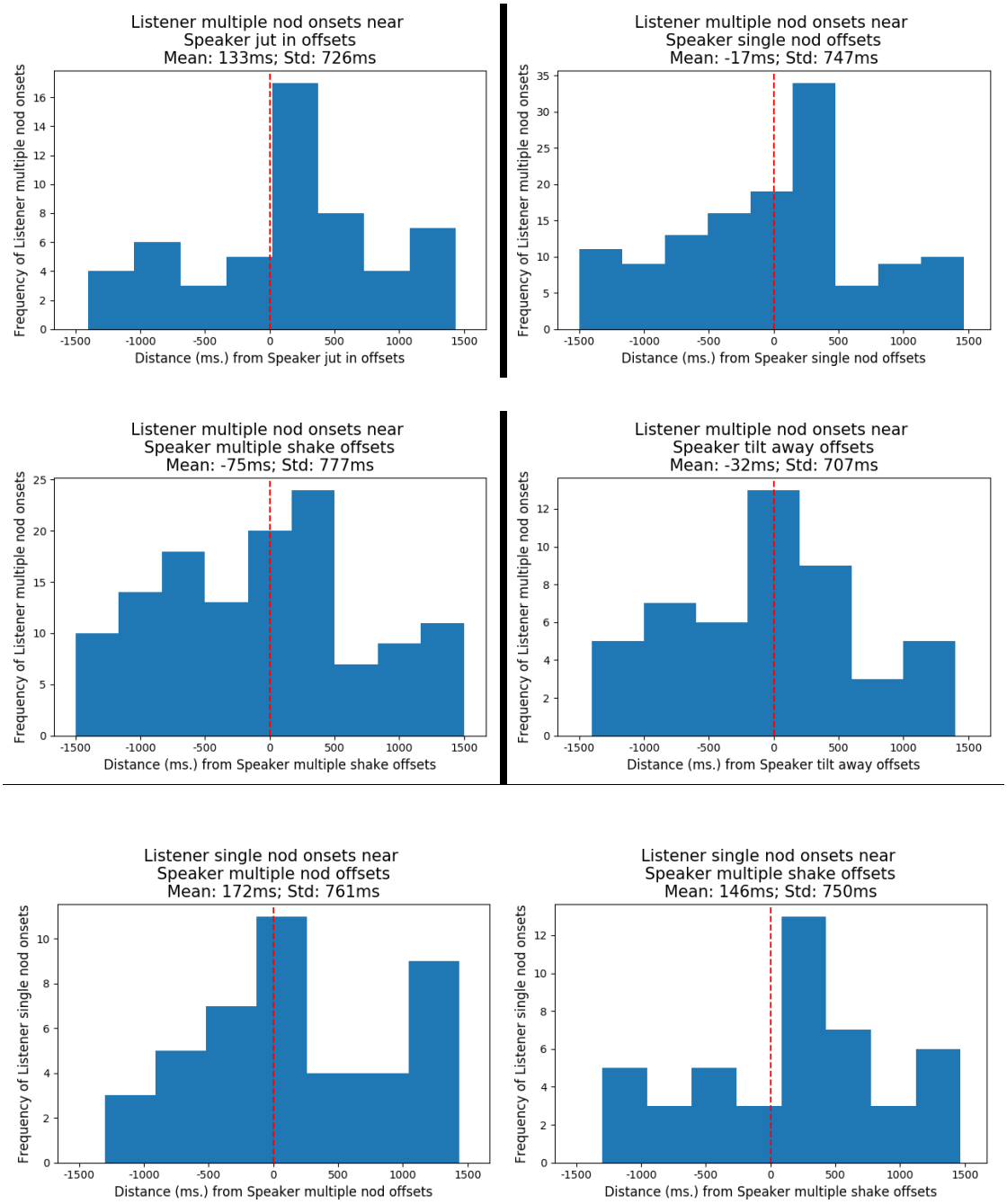
Figure 51. Window histogram – Speaker nod onsets near Listener head gesture onsets



Speaker head nods tend to peak around the offset of listener behaviors, whether nods up, retractions back, or tilts away, but they tend to precede the offsets of nods up, and follow the offsets of the retractions and tilts. Listener multiple nods are the most frequent listener head behaviors, but there was not clear pattern between the offsets of multiple nods and the onsets of speaker head behaviors, possibly because it can be more difficult to know when the end of a chain of nods will occur, and speakers have often looked away before the nodding is finished. (Note that the listener offsets here are all half-cycles, which tend to have an easily detectable endpoint.)

However, the onsets of multiple nods do pattern fairly closely with the offsets of some speaker head gestures, as do listener single nods.

Figure 52. Window histograms – Listener nod onsets near Speaker head gesture offsets



This timing generally takes of the form of increasing likelihood of a nod leading up to the onset of the speaker head gesture, with a peak in likelihood coming just after this offset. Many of these have fairly low token counts, but the timing pattern is still quite clear.

The fact that such a timing pattern exists at all is actually rather interesting. The fact that speech onsets are timed to follow speech offsets makes acoustic sense: since speakers and listeners share the same acoustic space, any overlap increases the difficulty in interpreting either individual signal. But for head gesture, and other non-verbal articulations, there is no shared space to clutter up – each participant has their own visual field. Another explanation for sequences of isolated behaviors (like speech turns) is that these articulations communicate meaning, and this meaning is first interpreted (perhaps as a whole unit), and then responded to in turn. Of course, an alternative explanation might be that some of these head gestures simply accompany speech, and the sequential nature of the head gestures reflects the sequential nature of speech (although most listener head gesture in this corpus is not co-speech).

4. Within-Mode Co-Occurrence Patterns: Speech Subtypes

4.1 Likelihood Measures

4.1.1 Speaker Speech Turns + Listener Speech Turns

We saw in Figure 45 that a substantial proportion of listener speech (and a very small proportion of speaker speech) overlaps with the interlocutor's speech. This overlap is distributed across different subtypes of speech, and these subtypes can differ substantially in the nature of this overlap. We will look first at how different listener turn types overlap with speaker turn types. Because there are so few tokens of listener fillers (N=5) and incompletes (N=18), these will be omitted from this analysis. Conditional probabilities of overlapping speech turn types are shown below as proportions of listener speech turns in Table 45, and of speaker speech turns in Table 46.

Table 45. Conditional probabilities - P(Speaker speech turns | Listener speech turns)

Listener speech type	Speaker speech type				
	Declarative	Interrogative	Filler	Incomplete	Back-channel
Listener declarative	**0.260**	0.004	0.003	0.020	0.091
Listener interrogative	**0.228**	0.021	0.017	0.031	0.018
Listener back-channel	**0.470**	0.083	0.015	0.039	0.012

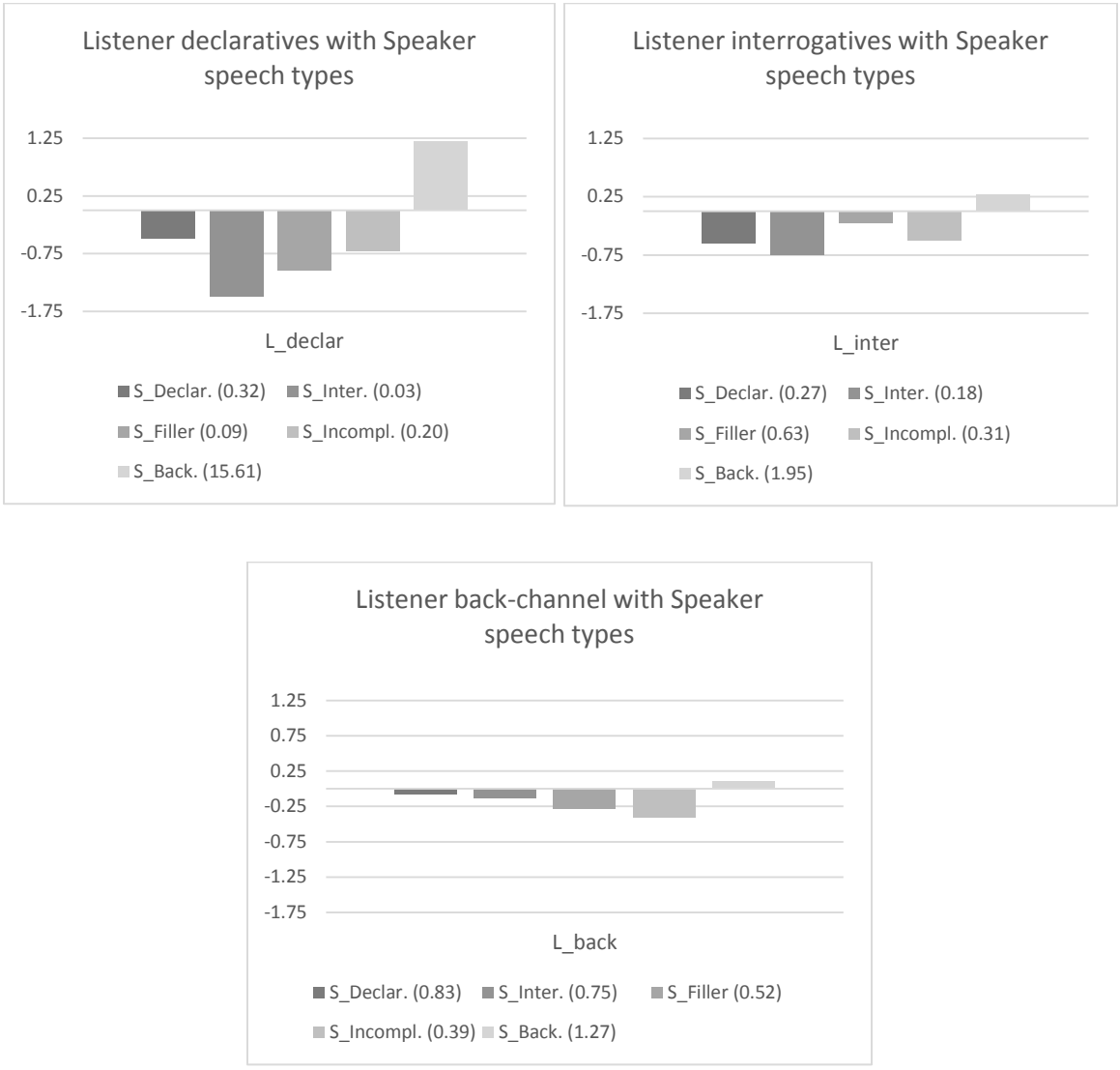
Table 46. Conditional probabilities - P(Listener speech turns | Speaker speech turns)

Speaker speech type	Listener speech type		
	Declarative	Interrogative	Back-channel
Speaker declarative	0.018	0.009	0.056
Speaker interrogative	0.001	0.004	0.048
Speaker filler	0.003	0.013	0.033
Speaker incomplete	0.008	0.007	0.026
Speaker back-channel	**0.346**	0.039	0.076

In Table 45, we see that speaker declaratives occupy the bulk of the speaker speech that overlaps with listener speech. They overlap with nearly half of listener back-channel frames (0.47), and with around a quarter of listener declarative (0.26) and interrogative turns (0.23). Other speaker speech types account for much less of this overlap, with speaker interrogatives (0.08) and incompletes (0.04) accounting for the next largest portions.

In Table 46, we see that listener declaratives are the only listener behavior to account for any substantial portion of overlap with speaker speech types, overlapping with more than a third of speaker back-channel frames (0.35). It would see that, across roles, back-channels overlapping with declarative speech are the most frequent form of overlap.

Figure 53. Odds ratios of observed overlaps of Listener speech types with Speaker speech types



Comparing across the three kinds of listener speech types (declaratives, interrogatives, and back-channels), we see in Figure 53 that the greatest amount of attraction and repellence occurs with listener declaratives. For attraction, they are 16 times more likely than expected to be overlapped with speaker back-channels (15.61). For repellence, they are one third as likely to overlap with speaker interrogatives (0.03), one eleventh as likely to overlap with speaker fillers (0.09), one fifth as likely to overlap with speaker

incompletes (0.20), and one third as likely to overlap with speaker declaratives (0.32). When listeners take declarative turns in this kind of storytelling context, they seem mostly likely to avoid overlapping with interrogatives, followed by speech that is incomplete, followed by declarative speech.

These differences from expected overlaps are much smaller in listener interrogatives (Figure 53 – interrogatives), but still quite significant. Speaker back-channels are only twice as likely as expected to overlap with listener interrogatives (1.95). All other speaker speech types are less likely than expected, but not to the same degree as for listener declaratives, although it seems that speaker interrogatives are still the least likely to overlap (0.18).

In Figure 53 – back-channels, we see even more reduced differences from a chance distribution. Listener back-channels are slightly more likely than expected to overlap with speaker back-channels (1.27). They are also only slightly less likely than expected to overlap with speaker declaratives (0.83), and this likelihood decreases for other speaker speech types. Listener back-channels are more likely to occur in the pauses between speech, but it seems like when they do overlap with speaker speech, it is more likely to be with speech that contains predication or is interrogative, and less likely during incomplete utterances. This is likely due to the fact that back-channels tend to be in response to completed utterances.

4.1.2 Speaker Speech Turns + Listener Back-Channels

Next, we turn to the co-occurrence patterns between speaker speech turn subtypes and listener back-channel subtypes. These are the categories that make up the majority of both speaker and listener speech, so the speech co-occurrence patterns found here are the

most frequent in the corpus. This means that there are sufficient tokens to show overlap data for each subtype for each role. Conditional probabilities are shown below in Table 47.

Table 47. Conditional probabilities - Speaker speech turns given Listener back-channels

Listener back-channel	**Speaker speech type**				
	Declarative	**Interrogative**	**Filler**	**Incomplete**	**Back-channel**
Acknowledgment	0.507	0.125	0.002	0.021	0.001
Assessment	0.472	0.047	0.014	0.047	0.012
Laugh	0.538	0.027	0.038	0.059	0.017
Continuer	0.485	0.118	0.015	0.015	0.000
Affirmation	0.479	0.106	0.006	0.003	0.015
Coll. Finish	0.453	0.098	0.000	0.108	0.049
Newsmarker	0.428	0.072	0.019	0.045	0.005

The first thing to note in Table 47 is that approximately half of all listener back-channels overlap with speaker declaratives. Some overlap more, like laughs (0.54), and some less, like newsmarkers (0.43), but none are strikingly dissimilar with respect to how much they overlap with declaratives. While listener back-channels overlap less with other speech behaviors, they show greater variance in how much they overlap. With speaker interrogatives, listener acknowledgments, continuers, affirmatives, and collaborative finishes all show an overlap of 10% or more, while assessments overlap less than 5% of the time, and laughs less than 3% of the time. Assessments and laughs both tend to be responses to new information, and interrogatives are less likely than some other speech turn subtypes to contain new information. Overlap proportions for most other behaviors tend to be fairly small, but we see that 11% of collaborative finish frames overlap with incompletes and 5% with speaker back-channels – these incompletes are often being

completed by the collaborative finish, and the speakers often respond to a collaborative finish with an affirmative back-channel.

How these overlaps differ from chance can be seen in Table 48.

Table 48. Odds ratios of observed overlaps (of speaker speech turns with listener back-channels) to expected overlaps

Listener back-channel	**Speaker speech type**				
	Declarative	**Interrogative**	**Filler**	**Incomplete**	**Back-channel**
Acknowledgment	ns.	**1.20****	0.08***	0.21***	0.12***
Assessment	0.84*	0.40***	0.51***	0.48***	ns
Laugh	**1.11***	0.23***	**1.43*****	0.62***	**1.89*****
Continuer	ns.	ns.	0.56*	0.15***	0.00***
Affirmation	ns.	ns.	0.22***	0.03***	ns.
Coll. Finish	0.78**	ns.	0.00***	ns.	**5.52*****
Newsmarker	0.71**	0.65*	ns.	0.47*	ns

The majority of pairs of behaviors overlap less than expected, or their overlap is not significantly different from the expected overlap, which is in keeping with the window of opportunity concept. However, some behaviors do overlap more than would be expected. For the most frequent kind of speaker behavior, declaratives, listener laughs are more frequent than expected, while assessments, collaborative finishes, and newsmarkers are less frequent. As mentioned above, assessments and laughs tend to be responses to new information, but it seems that laughs are more likely than assessments to overlap with speaker declaratives.

Listener acknowledgments are twenty percent more likely to overlap with speaker interrogatives than expected. These are typically in response to rising intonation (acknowledging what has been said, but not directly affirming it), and this overlap usually immediately precedes the end of the speaker interrogative. On the other hand,

assessments, laughs and newsmarkers are much less likely, in keeping with the theory that these are responses to new information. Interestingly, listener affirmatives are at chance with respect to speaker interrogatives, but this may be because listeners wait until a yes-no question has been completed to speak, or because there are so few direct questions in the corpus.

Most listener back-channels are much less likely to occur than expected during speaker fillers and incomplete predications, although this is truer for some back-channels than others. Acknowledgments are less than one tenth as likely during fillers, while continuers and assessments are not quite half as likely as expected. During incompletes, nothing is more likely than expected to overlap. During speaker back-channels, continuers and acknowledgments are much less likely, but assessments are at chance, while collaborative finishes are much more likely than expected, as speakers typically respond to a collaborative finish with an affirmative, or a reciprocal collaborative finish.

Laughs are the listener behavior most likely to have a greater degree of overlap than expected, being more likely during speaker declaratives, fillers, and back-channels. It may be that laughter, compared to other vocal behaviors, interferes less with speech comprehension because it contains no language, and is therefore an acceptable auditory behavior during others' speech.

4.2 N-grams

To look at the sequential patterns of speech segments across roles, we will examine the bigram frequencies of speaker and listener speech-turn boundaries that occur within one second of each other, ordered by their symmetrical conditional probability (Table 49). Only bigrams with a symmetric conditional probability of 0.05 or higher were included.

Table 49. Listener and Speaker Speech bigrams (1-second window)

Bigram	Frequency	Symm. CP	CP: 1|2	CP: 2|1
S. declarative offsets + L. back-channel onsets	132	0.030	0.232	0.129
L. interrogative offsets + **S. back-channel onsets**	12	0.020	0.124	0.160
S. interrogative offsets + L. back-channel onsets	48	0.020	0.085	0.231
L. back-channel offsets + **S. declarative offsets**	104	0.019	0.102	0.188
L. declarative offsets + **S. back-channel onsets**	12	0.017	0.124	0.138
L. back-channel offsets + **S. declarative onsets**	81	0.012	0.079	0.146
S. declarative onsets + L. back-channel offsets	62	0.007	0.112	0.060
L. back-channel offsets + **S. filler onsets**	30	0.006	0.117	0.054
L. interrogative offsets + **S. declarative onsets**	22	0.006	0.021	0.293
L. back-channel offsets + **S. incomplete onsets**	30	0.005	0.089	0.054
L. back-channel offsets + **S. interrogative offsets**	23	0.005	0.111	0.042

It is not a surprise that most of the most dependent bigrams involve a speech onset following a speech offset or (the three exceptions all involve speaker offsets following listener back-channel offsets or speaker onsets preceding listener back-channel offsets, all instances of speakers' turns intruding on listener back-channels). Of the offset + onset pairs, speakers and listeners seem evenly mixed, in terms of who appears to be initiating the sequence (although we saw in Section 2.3.3 that speakers do not time their onsets to interlocutor offsets as precisely as listeners do). Of the instances where speaker offsets initiate the sequence, we see only declarative and interrogative speech segments, and no incompletes, fillers, or speaker back-channels. Likewise, we only see speaker back-channels responding to listener declaratives and interrogatives (although there are very few listener incompletes or fillers).

We can look in more detail at how different subtypes of listener back-channels are associated with different speaker speech types (Table 50).

Table 50. Speaker speech turn and Listener back-channel bigrams (1-second window)

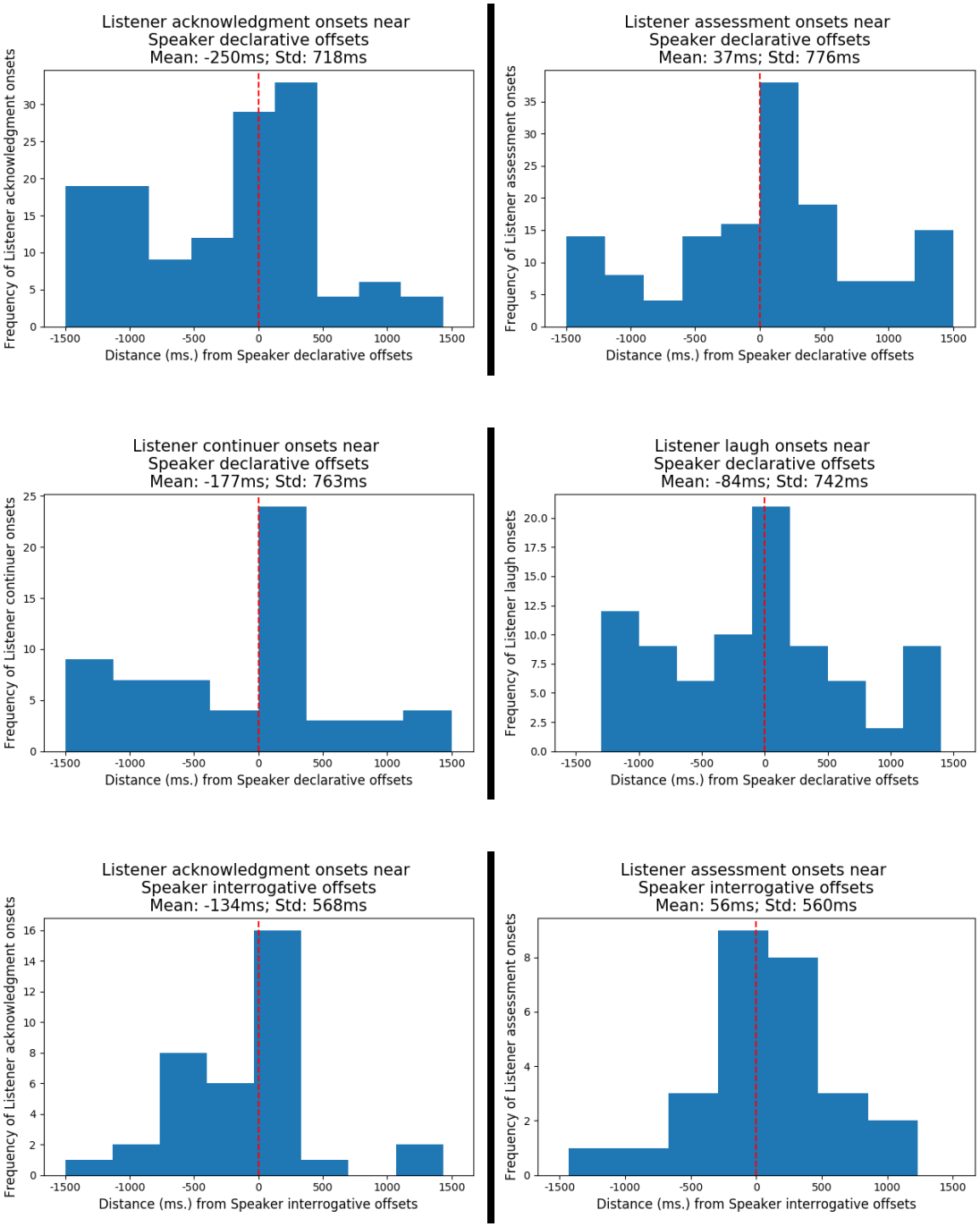
Type	Bigram	Freq.	Symm. CP	CP: 1|2	CP: 2|1
S-off/L-on	**S. declarative offsets** + L. acknowledgment onsets	48	0.011	0.241	0.047
S-off/L-on	**S. declarative offsets** + L. assessment onsets	49	0.011	0.230	0.048
S-off/L-on	**S. interrogative offsets** + L. continuer onsets	12	0.008	0.136	0.058
L-off/S-off	L. acknowledgment offsets + **S. declarative offsets**	39	0.008	0.038	0.205
L-off/S-off	L. assessment offsets + **S. declarative offsets**	34	0.006	0.033	0.166
S-on/L-off	**S. declarative onsets** + L. assessment offsets	34	0.005	0.166	0.033
S-on/L-on	**S. declarative offsets** + L. continuer onsets	21	0.005	0.239	0.021
S-off/L-on	**S. interrogative offsets** + L. acknowledgment onsets	14	0.005	0.070	0.067
L-off/S-on	L. acknowledgment offsets + **S. declarative onsets**	30	0.005	0.029	0.158
L-off/S-off	L. continuer offsets + **S. declarative offsets**	20	0.004	0.020	0.222
L-on/S-off	L. laugh onsets + **S. declarative offsets**	21	0.004	0.021	0.202
L-off/S-on	L. acknowledgment offsets + **S. filler onsets**	14	0.004	0.054	0.074
LSoffon	L. laugh offsets + **S. declarative onsets**	22	0.004	0.021	0.188
LSoffon	L. assessment offsets + **S. declarative onsets**	29	0.004	0.028	0.141
SLoffon	**S. interrogative offsets** + L. assessment onsets	13	0.004	0.061	0.063

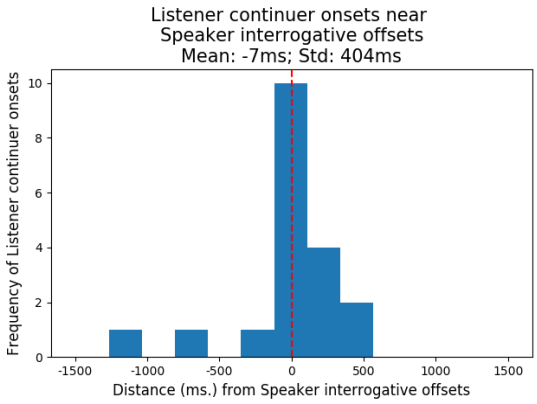
Listener acknowledgments and assessments onsets are strongly linked to speaker declarative offsets. They can occur just before or just after, but are more likely to occur just after. Listener continuers are more closely tied to speaker interrogative offsets (and also to declarative offsets with a slightly weaker dependence), and these are less likely than acknowledgments and assessments to precede the speaker speech offset. This may seem unintuitive, given that continuers are the kind of back-channel that responds least to the content of the speaker message, as opposed to acknowledgments and assessments, but it is possible that continuers are in fact responding more to the intonation of the speaker speech, and the primary feature of interrogative prosody, the final rising pitch, often occurs at the very end of a speech segment.

4.3 Window Histograms

Despite the fact that many of the speech subtypes do not have very numerous tokens, there are still a number of window histograms that show fairly clear timing patterns. We will look first at listener back-channel onsets occurring near speaker speech offsets.

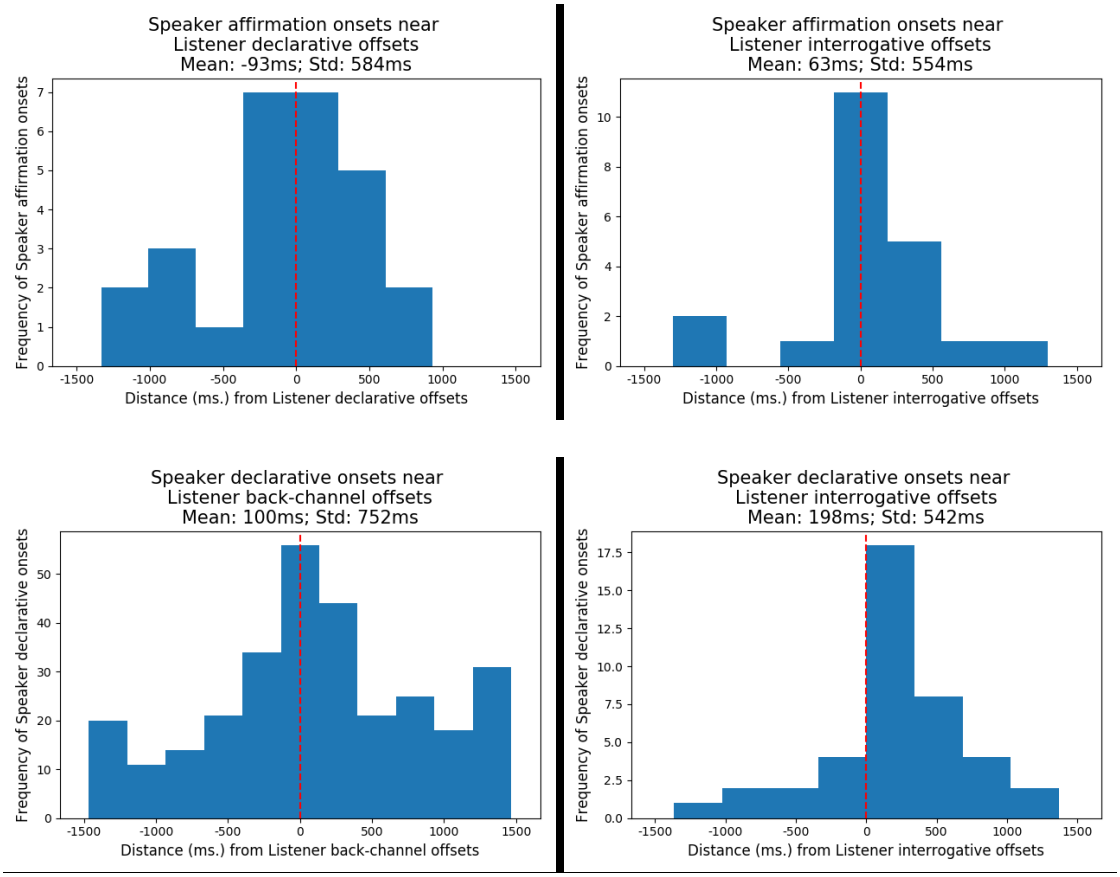
Figure 54. Listener back-channel onsets near Speaker declarative and interrogative offsets





As we can see in the figures above, all listener back-channels are timed quite precisely to meet the offset of the preceding speaker speech, whether it is declarative or interrogative. For some back-channels, there is a greater chance of the onset also preceding the speaker offset, as with most speaker declaratives, and acknowledgments near interrogatives. Perhaps the most precisely timed is the listener continuer onset near speaker interrogative offsets – over half of these tokens occur within a few frames of the speaker offset. As mentioned in Section 4.2, this may be a result of continuers responding more to the intonational contour of the speech than the content.

Figure 55. Window histogram – Speaker speech onsets near Listener speech offsets



For speakers, affirmations are the only back-channel with sufficient tokens to examine in detail. Like listener continuers, they are more precisely timed to co-occur with interrogative offsets than declarative offsets. This may also be due to the acoustic nature of interrogatives, or it may speakers are more willing to interrupt a listener declarative (because it is *their* time to talk, not the listener's) than an interrogative.

5. Summary and Hypotheses

5.1 Summary

In this chapter, we looked at the timing relations between behaviors of the same modality, across speakers and listeners. In Section 2, we looked at patterns of mutual overlap across roles, seeing that periods of mutual overlap were relatively infrequent for most

modalities, except gaze, where mutual gaze accounted for 27% of the total corpus. Listener-only and speaker-only proportions were small in most modalities, being most similar in the head modality. It was found that overlapping heads were much more likely than would be expected from chance, that mutual gaze was very close to chance, and that speech and manual gesture overlap was much less than would be expected from chance.

We also saw that the most frequent sequences of gaze shift involved a ‘chasing’ pattern, as in the speaker looking at the listener, the listener looking away, the speaker looking away, and the listener looking back at the speaker. Gaze seems to repel other gaze after a period of time.

Looking at window histograms of each modality, we saw some timing relationships between onsets and onsets of modalities, but these were not nearly as clear as onsets following offsets. Onsets following offsets was extremely clear in the speech modality, where listeners showed even more precise timing than speakers, suggesting that in this kind of communicative context, there is more focus on allowing the speaker to speak. There was also a timing relationship between head offsets and head onsets, both for speakers and listeners. These were not timed with the same precision as speech, likely in part because head gestures do not impede each other like speech does, but speakers and listeners were also more similar in the precision of their timing across roles.

In Section 3, we looked in more detail at the timing relations between subtypes of head gestures across roles. Several subtypes were found to co-occur across roles more often than expected from chance, including multiple nods, single nods, nods up, and tilts away + return. Other subtypes were found to have a greater likelihood to co-occur with other interlocutor head behaviors: speaker retractions back were more likely to occur with most

listener tilts and juts, while speaker juts in were found to be less likely to occur with these same listener behaviors. Speaker shakes, which are often co-produced with negative affect speech content, were more likely to co-occur with listener juts in and retractions, which often express horror or amazement.

In Section 4, we looked in more detail at the timing relations between subtypes of speech across roles. The only speech subtypes more likely than expected to overlap were speaker back-channels, which were slightly more likely to overlap with listener back-channels, and substantially more likely to overlap with listener declaratives. Looking at listener back-channels, we found that listener acknowledgments and assessments were more closely tied to speaker declaratives, and that listener continuers were more closely tied to speaker interrogatives.

5.2 Hypotheses

There are a number of hypotheses one could formulate from the data in this chapter. A small sample is laid out below.

1. Some theorize that speech production is designed to distribute information optimally over time in a uniform manner (Bell et al. 2003, Aylett & Turk 2004). An alternative theory might be that speakers distribute information with varying degrees of density, for pragmatic reasons. These theories could be tested by examining the amount of information being co-produced on a variety of modalities, including non-verbal modalities such as those in this corpus as well as intra-linguistic modalities.
    - If communicative information is distributed uniformly over time, we might hypothesize that a speaker's multimodal information will co-occur with

speech information that is unpredictable or otherwise more difficult to process. For example, iconic manual gesture might co-occur with speech to provide redundant information that is hard to process, or to provide additional semantic information.

2. Are other mutual modality segments (such as co-occurring head or manual gesture) also more likely during mutual gaze?
3. Does speech content during mutual gaze (or mutual head gesture) differ qualitatively from speech content during other gaze combinations?
    - We might hypothesize that speech during mutual gaze is, in some way, more 'important,' and so it is produced while the speaker can assess its effect on the listener. This might mean that the information is more salient to the narrative, more surprising, or more calculated to have an emotional impact. These are rather subjective categories, but different passages of the narrative could be subjectively rated on these scales by participants reading the text. If the speech segments that occur during mutual gaze are rated higher on these categories, this would support this hypothesis.
4. Where do 'enclosed' back-channels (back-channel speech or head gesture that begins after a speaker speech onset and ends before the speech offset) occur within the speech segment? Are they responding to different kinds of information than back-channels that occur at the end of speech segments?
    - Back-channels often occur following a speaker speech segment, but also often occur during speaker speech. One hypothesis that would be relatively simple to test is that back-channels that occur during speaker speech are more likely

than not to occur during *longer* speaker speech segments (either because there is more information to respond to, or because the listener grows tired of waiting for an opening – if it is the latter these might tend to occur more towards the end of long segments).

- Another hypothesis might be that back-channels respond to speech information that may or may not occur at the end of speech segments. For example, some kinds of information that back-channels may respond to are new information, event completions, or uncertainty. If back-channels are more likely than expected following this kind of information when it is both at the end of a speech segment and internal to the speech segment, this would be good evidence to support this hypothesis.

5. We have seen that continuers are more closely tied to interrogatives than declaratives, and suggested that their precise timing relative to interrogative offsets may be because they are responding more to intonation than speech content. Do back-channels respond more to semantic or acoustic information?
    - According to one hypothesis, back-channels might respond more to acoustic information, while according to the other they would respond more to semantic information.
    - To manipulate the amount of semantic information, we could run two groups: in one a listener must respond to a storyteller whose speech is garbled, in another language, or in some way unintelligible, and in the other the listener responds to a story they understand. If the listener's back-channels

(particularly continuers) exhibit the same timing characteristics across groups, this would support the first hypothesis.

- To manipulate the amount of acoustic information, we could have two groups: in one a listener must respond to a storyteller whose pitch has been flattened, and in the other the listener responds to normal storytelling. If the timing characteristics of the back-channels are similar across groups, this would support the second hypothesis

6. Speaker and listener head gesture onsets are more closely tied to interlocutor head offsets than interlocutor head onsets.
    - One hypothesis might hold that participants time the onsets to the offsets, and are able to predict the offset based on the conventional shapes these gestures usually take. Another hypothesis might hold that they time onsets to onsets, and the pattern we see is because the time it takes to plan and produce their head gesture response is similar to the average duration of the head gesture they are responding to.
    - The first hypothesis could be explored by looking at the timing of head gesture onsets to the offsets of head gestures that are nonstandard (and thus unpredictable) in some way, either in shape or duration or some other feature. The second hypothesis could be explored by looking at the variance in the lag between the onset of the first head gesture and the onset of the second – if true, this should be small.

CHAPTER VII: ACROSS-ROLE / ACROSS-MODALITY

1. Introduction

This chapter examines the timing relationships between behaviors produced by different people, in different modalities. There is less reason to think that these kinds of relationships will pattern together, compared to relationships in previous chapters. Within-role, across-modality relationships (Chapter 5) are produced by the same motor system, and are theorized to arise from the same 'growth point' and have the same communicative goal. Across-role, within-modality relationships (Chapter 6) share the same articulators, which are the most natural candidates to imitate and/or respond to each other, assuming they communicate the same kinds of information. These are reasons to predict that there will be synchronization and co-production in these kinds of relationships. For across-role, across-modality relationships, these motivations are absent (although indirect timing patterns may appear because of interactions in the other two kinds of relationships – e.g. if listener speech occurs following speaker speech, and listener head gestures co-occur with listener speech, we would find that listener head gestures also occur following speaker speech), but there may be other motivations for some of these modalities to pattern together, which may be obvious or not obvious. Of course, there will also be many areas where we would expect no pattern, and we find no pattern, such as between speaker head gestures and listener manual gesture. No one has reported an interaction between these role-modality combinations, and no one has suggested a motivation for such an interaction existing. But in this way, this chapter is the perfect place for this dissertation's look-at-everything approach to multimodal analysis, because it may allow us to identify relationships in places where we had no reason to expect them.

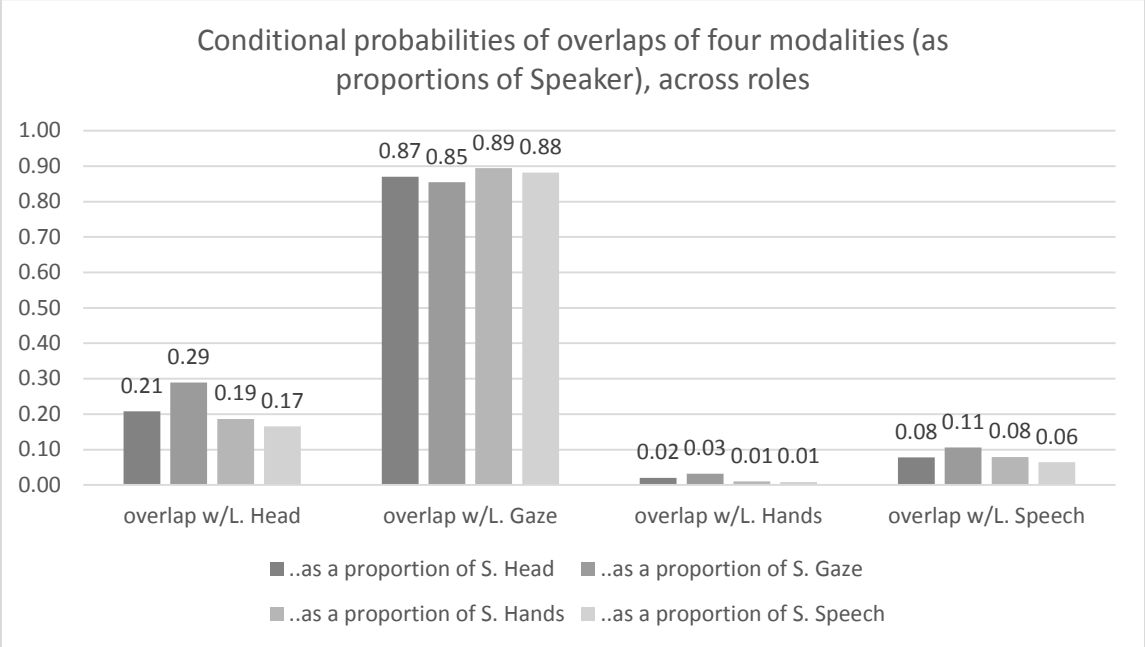
The chapter is divided like the previous chapters. Section 2 looks at all four modalities broadly, using likelihood measures, n-grams, and window histograms. Section 3 looks at timing relationships between subtypes of head and speech behavior. Section 4 looks at the timing of head and gaze behavior, an area where we find a great deal of interaction. Section 5 examines timing relationships between gaze and speech. In Section 6, we look at the timing between gaze and manual gesture. Finally, Section 7 offers a summary of the findings, and a set of hypotheses based on the findings.

2. Overview of the Four Modalities

2.1 Likelihood Measures

We begin by looking at all four of the modalities, undivided by subtypes. The figures below show the conditional probabilities of the overlaps between each pair of modalities, first as proportions of the speaker behaviors (Figure 56), then as proportions of the listener behaviors (Figure 57). For comparison's sake, within-modality pairs are also included.

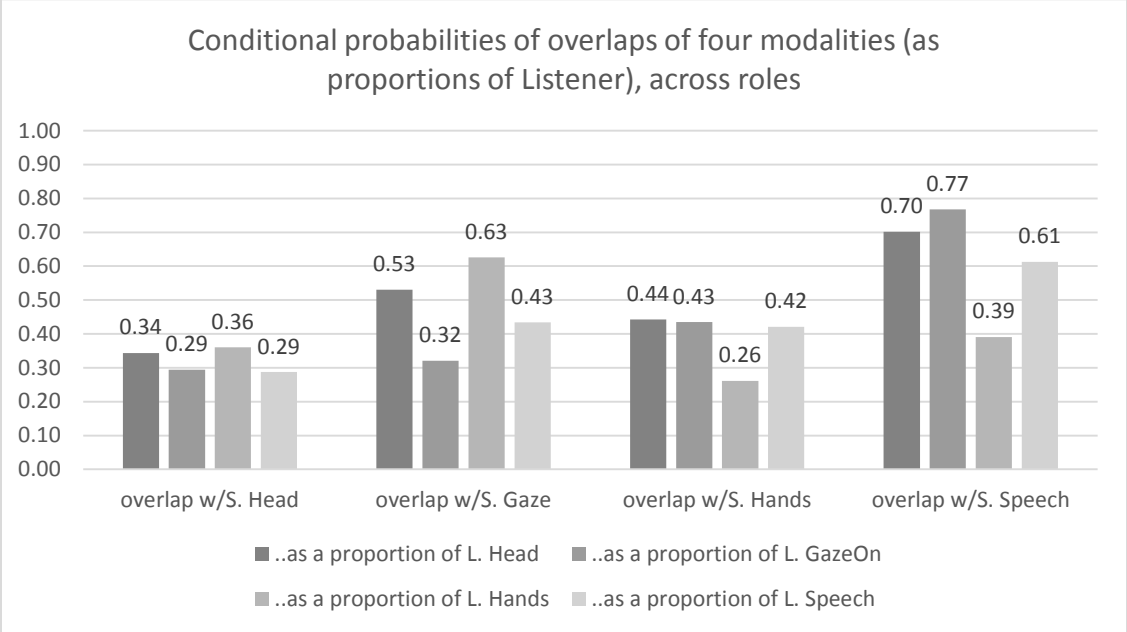
Figure 56. Conditional probabilities of overlaps of four modalities (as proportions of Speaker), across roles



The first thing to stand out is that listener gaze overlaps with the vast majority of all speaker behaviors (including speaker gaze-towards). Listeners gaze at speakers for around 85% of the entire corpus, and so the fact that it overlaps with all these behaviors for at least that proportion of time suggests that these speaker behaviors may be timed to co-occur with it.

Other listener behaviors overlap much less with speaker behaviors, although listener head gesture overlaps more than speech or manual gesture. Also, these other three listener modalities show a similar pattern in terms of the ordering of how much of each speaker behavior they overlap with. Listener heads, hands, and speech each show the greatest overlap proportion with speaker gaze-towards, followed by speaker head and hands, and then by speaker speech.

Figure 57. Conditional probabilities of overlaps of four modalities (as proportions of Listener), across roles



Conditional probabilities for proportions of listener behaviors (Figure 57) show more variability than for of speaker behaviors. Speaker speech accounts for the largest proportions of listener behavior, particularly for listener head and listener gaze, but also for listener speech (which we saw in Chapter 6), but this drops to only 39% for listener hands, which occur more with listener speech turns than listener back-channels.

Speaker heads account for the smallest overlap, clustering around one third of the duration of listener behaviors. Speaker hands account for more overlap: a little over 40% for all listener behaviors except listener hands, which overlaps the least of all modalities. Speaker gaze overlaps the most with visible listener behaviors, accounting for over 50% of listener head and manual gesture duration (even more with manual gesture), only 43% of listener speech, and only 32% of listener gaze.

To look at how these overlaps compare to overlaps predicted based on their overall frequencies, we look at their odds-ratios in Figures 58, broken down by listener behaviors.

Figure 58. Odds ratios (log-transformed) of Listener head and gaze-towards with Speaker behaviors

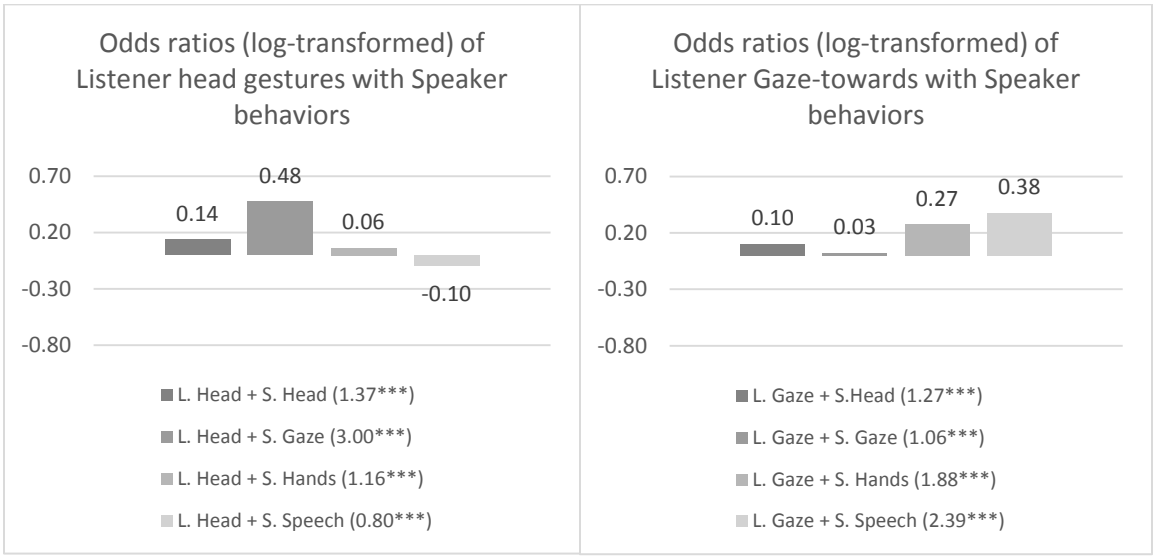


Figure 58a: Listener head gesture. As seen in Chapter 6, listener heads and speaker heads overlap more than expected. To a lesser extent, so do listener heads and speaker hands, although this is not much different from chance. They are especially likely to overlap with speaker gaze, though, being visible behaviors. Listener heads are less likely than expected to overlap with speaker speech – the certainly do overlap, but they seem to occur relatively more often during speaker pauses.

Figure 58b: Listener gaze. This is the only listener behavior that overlaps more than expected with *all* speaker behaviors (although only barely with speaker gaze). The difference from expected is not nearly as great here for speaker heads as it was for listener heads, but listeners gaze-towards is much more frequent. Listener gaze during

speaker manual gesture is even more likely than during head gesture, but even this is not as likely as during speaker speech. The difference between speaker gaze + listener heads and listener gaze + speaker heads is quite substantial, suggesting that listener head gestures may be more designed to be seen (and thus more communicative) and speaker head gestures may be less so (and thus more for facilitation of speech production). The fact that speaker speech is the greatest draw for listener gaze is also interesting, given that speech is not a visible behavior, but there are at least two possible motivations for this: 1) gaze is itself communicative, signaling attention when the speaker is producing their message, and 2) there are a number of other modalities in the face that interact with speech that have not been coded, including facial expressions and blinks.

Figure 59. Odds ratios (log-transformed) of Listener manual gesture and speech with Speaker behaviors

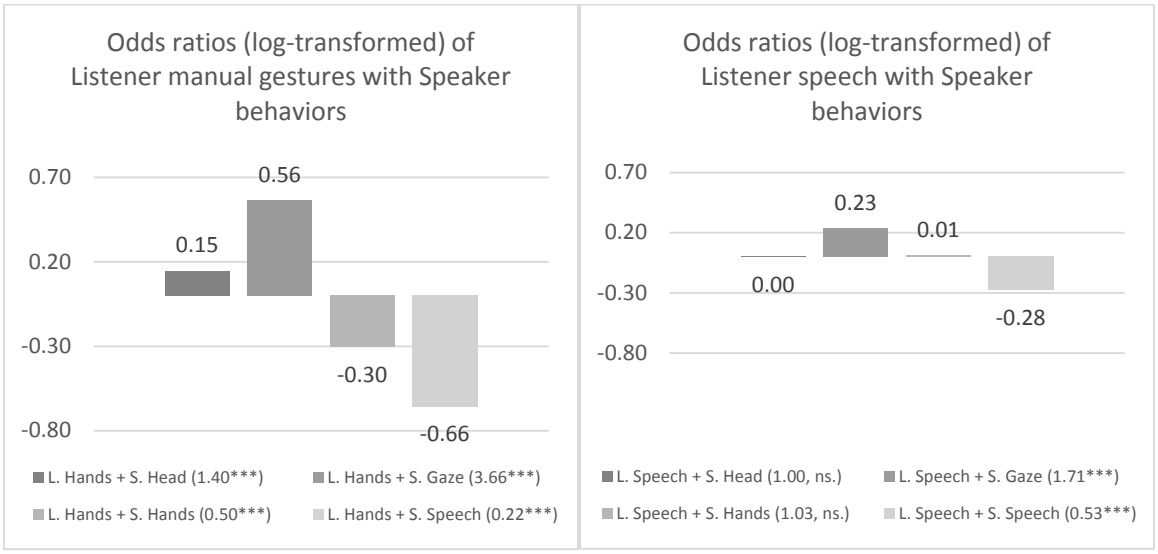


Figure 59a: Listener manual gesture. Listener manual gesture is even more likely to overlap with speaker gaze (3.66) than speaker manual gesture with listener gaze (1.88). It is also more likely than expected to overlap with speaker heads (speakers are nodding in affirmation during listener turns, and often head-gesturing during back-channels). They

are much less likely than expected to overlap with speaker hands and speech, because manual gesture is co-speech in both roles, and speech does not overlap a great deal across roles.

Figure 59b: Listener speech. Listener speech is at chance with speaker head manual gesture. There seem to be no dependencies between these behaviors. It is more likely than expected during speaker gaze – listener speech turns and back-channels are collapsed here, so this will be influenced by the fact that speakers tend to look at listeners when listener take speech turns. Of course, listener speech is also less likely than expected during speaker speech, because of the nature of turn-taking. Some listener back-channels do overlap more with speaker speech, as we will see in Section 3, but listener speech turns tend not to (and listener co-speech manual gesture is more common with listener speech-turns, explaining why the overlap between listener hands and speaker speech is less likely the overlap of listener speech and speaker speech).

Several of the more interesting findings in these figures have had to do with which behaviors are more or less likely to occur during interlocutor gaze-towards. Most behaviors are more likely to be produced when the producer is being looked at, but to different degrees across modalities and roles. For example, we have seen how speaker speech and manual and head gesture are more likely during listener gaze-towards, but speaker head gesture is less likely than speech, despite it being a visible behavior. We also know, from Figure 26 in Chapter 5, Section 2, that speaker head gesture is more likely to be produced during speaker gaze-towards. One way to get a better idea of the relationship between gaze and head gesture would be to look at how frequently it is produced when both speaker and listener are looking at each other, during what is called

mutual gaze, or mutual gaze-towards. It might be equally interesting to look at co-occurrence patterns during other logical combinations of gaze-towards and gaze-away, such as when only listeners or speaker are gazing-towards, or when both participants are gazing away, which is called mutual gaze-away here (see Figure 44 in Chapter 6, Section 2, for these distributions). Table 51 shows the odds ratios and conditional probabilities for these gaze combinations with each speaker and listener modality.

Table 51. Conditional probabilities and odds ratios of four modalities with gaze combinations

Gaze type	**Listener modalities**				**Speaker modalities**			
Odds-ratios of overlap								
	Head	**Gaze**	**Hands**	**Speech**	**Head**	**Gaze**	**Hands**	**Speech**
Mutual gaze-towards	**2.603*****	inf.	**1.715*****	**1.472*****	**1.749*****	inf.	**1.489*****	**1.438*****
Listener gaze only	0.337***	inf.	0.225***	0.474***	0.704***	0	ns.	**1.254*****
Speaker gaze only	**2.347*****	0	**6.886*****	**2.087*****	0.882***	inf.	0.388***	0.255***
Mutual gaze-away	**1.097*****	0	**1.348*****	**1.713*****	0.765***	0	0.645***	0.598***
Conditional probabilities of four modalities | Gaze overlap combinations								
	Head	**Gaze**	**Hands**	**Speech**	**Head**	**Gaze**	**Hands**	**Speech**
Mutual gaze-towards	0.285	1	0.023	0.099	0.375	1	0.484	0.788
Listener gaze only	0.108	1	0.007	0.055	0.257	0	0.412	0.758
Speaker gaze only	0.320	0	0.083	0.144	0.264	1	0.221	0.439
Mutual gaze-away	0.186	0	0.021	0.119	0.241	0	0.322	0.643
Conditional probabilities of gaze overlap combinations | Four modalities								
	Head	**Gaze**	**Hands**	**Speech**	**Head**	**Gaze**	**Hands**	**Speech**
Mutual gaze-towards	0.446	0.321	0.389	0.348	0.355	0.854	0.319	0.290
Listener gaze only	0.358	0.679	0.239	0.407	0.516	0	0.575	0.592
Speaker gaze only	0.085	0	0.237	0.086	0.043	0.146	0.025	0.028
Mutual gaze-away	0.112	0	0.135	0.159	0.087	0	0.081	0.091

The first thing to note is that, during mutual gaze (which makes up 27% of the corpus), every modality is more likely than would be expected from chance. However, these odds-ratios are lower for almost modalities than when looking broadly at listener and speaker gaze-towards. The two exceptions to these diminished odds-ratios are interlocutor gaze (which was close to being at chance in Figure 58b, but of course is entirely mutual here) and speaker heads, whose odds ratio increases from 1.27 with listener gaze-towards to 1.75 with mutual gaze-towards. During mutual gaze, speaker heads are now even more likely than speaker manual gesture and speech, which were more likely than speaker heads during listener gaze-towards (Figure 58b).

When we compare this to speaker behaviors when only listeners are gazing towards and speakers are gazing away (57% of the corpus), we see that speaker behaviors all become less likely. Speaker speech does remain more likely than expected, but speaker manual gesture is only at chance, and speaker head gesture is actually less likely than expected (0.70). Removing speaker gaze-towards creates a substantial and uneven difference in the likelihood of speaker behaviors. We suggested earlier that the reason speaker head gesture was less likely to overlap with listener gaze than speaker speech was that it had more to do with facilitation of speech production than communication of meaning, but this difference suggests the opposite. It suggests that speaker head gesture may often be communicative, precisely because it is more likely to be produced when the speaker can be seen, and can check for a response. In fact, mutual gaze is the only time when speaker heads are more common than expected, and this is true for speaker manual gesture as well.

For other listener modalities during listener-only gaze, everything is much less likely. This gaze combination is the most frequent, and the speech and manual and head gesture that are often co-produced during listener turns typically also involve gaze-away. This can be seen in the likelihoods of listener behaviors during speaker-only gaze (only 5% of the corpus), where listener heads remain highly likely, and listener manual gesture and speech become even more likely. During speaker-only gaze, speaker behaviors are all less likely than expected, especially speaker manual gesture and speech – this is probably influenced in part by listener turns, as well.

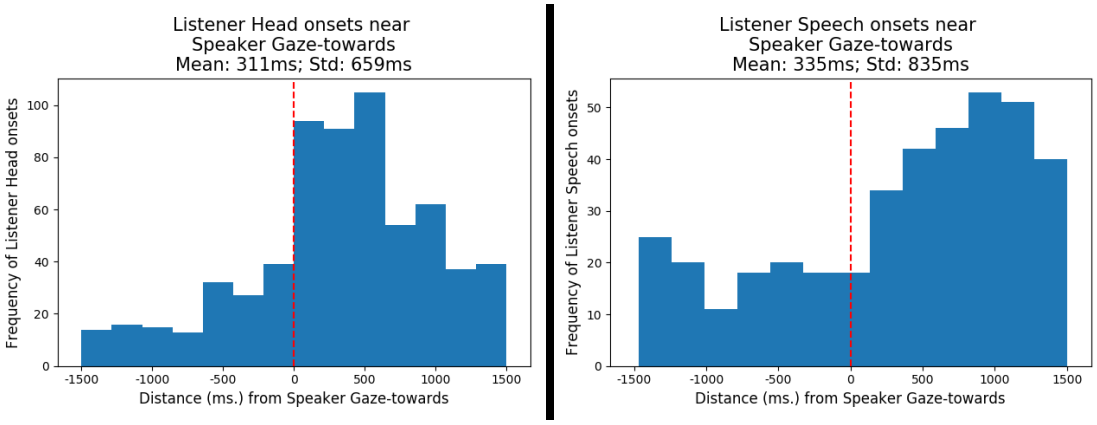
Finally, we see another asymmetry between roles in mutual gaze-away (10% of the corpus). When neither participant is looking at the other, speaker behaviors are all less likely than expected. On the other hand, listener behaviors are all more likely than expected (with the possible exception of listener head gestures – these are significantly more likely than expected (1.1), but this is not much higher than chance). Periods of mutual gaze-away tend to occur either during unfilled pauses in speaker speech (speakers rarely stare at listeners while they are trying to think of what to say next) or during listener responses, in cases where speakers don't take the opportunity to gaze towards the listener.

2.2 Window Histograms – Four Modalities

We now turn to an examination of the temporal relationships between onsets and offsets of the four modalities, without distinguishing between subtypes of head and speech behaviors. In this section, and even more so in later sections, there may be too few tokens to see patterns, and sometimes there are no clear patterns even when there are sufficient tokens. For these reasons, only selected histograms will be shown.

Beginning with onsets of speaker behaviors near the onsets of listener behaviors (onsets include gaze-shifts towards), the following figures show the frequency distribution of these behavior boundaries when they occur within 1500ms of each other.

Figure 60. Window histogram – Listener head and speech onsets near Speaker gaze-towards



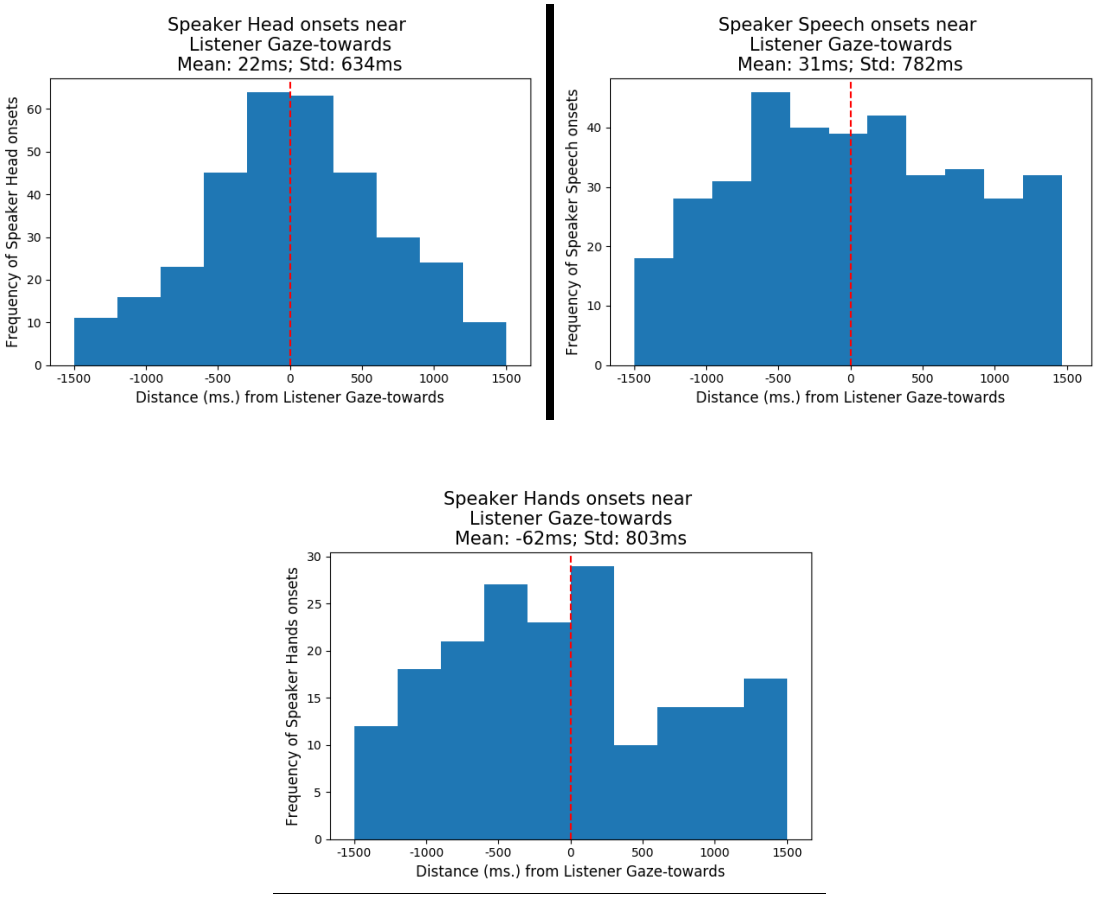
We see clear patterns of listener head and speech onsets relative to speaker gaze-towards. (There are too few tokens of listener manual gesture to see a pattern.) These onsets tend to be timed to occur just after the speaker's eyes lay on the listener. The timing of listener head gestures is particularly remarkable, with a sharp increase in frequency *precisely* at the moment of gaze-shift. Listener speech onsets are delayed slightly relative to head gesture onsets, in keeping with the finding from Chapter 5 that listener co-speech head gesture tends to precede the speech.

This is an across-role, across-modality timing relationship that is easily motivated. Speakers want indications of attention or comprehension, and listeners want to give them. Head gestures, the most common form of listener behavior (60% more frequent than listener speech segments) are only communicative when they are seen, so listeners have good reason to time their head gestures relative to this gaze-shift. Co-speech non-verbal

behaviors are typically analyzed as being produced to support speech, but this may be a case where the speech follows the gesture.

Of course, motor signals to produce head gestures do not travel instantaneously. There are clearly cues in the speaker's production that allow listeners to predict when they need to produce their head gestures: gaze shift rarely occurs without some sort of head turn, or other repositioning head gesture, and occasionally also a shift in the shoulders or the entire body.

Figure 61. Window histogram – Speaker head, speech, and manual gesture onsets near Listener gaze-towards

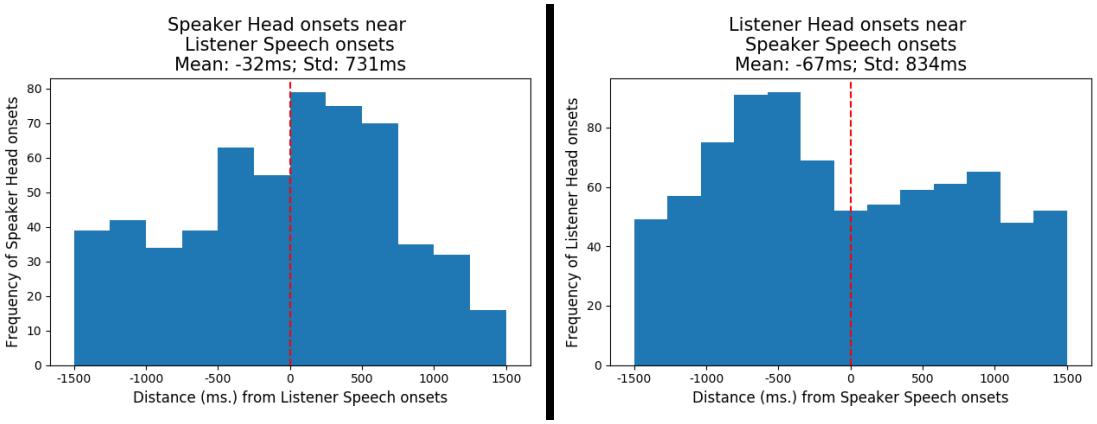


Looking at the timing relationships between listener gaze-towards and speaker onsets, we see a different pattern. First, while speaker head onsets do seem to be timed to co-occur

near listener gaze-towards, they don't show the same precision in their timing, being equally likely to occur before or after the gaze-shift. Still, this is more similar to listener onsets than we see in speaker speech onsets, which seems to have no clear timing to listener gaze-towards. Of course, speaker speech and listener gaze are both highly frequent, and speaker speech and speaker head gesture are not co-produced in the same way that they are for listeners.

Also interesting is the timing relationship between listener gaze-toward and speaker manual gesture onsets. Rather than manual gesture onsets becoming more frequent following listener gaze-towards, we see the reverse, with speaker manual gesture being more frequent before listener gaze. This suggests that speaker manual gesture actually has the effect of attracting listener gaze (and, to clarify, this is not gaze towards the hands, but gaze towards the face).

Figure 62. Window histograms – Speaker and Listener head onsets near Listener and Speaker speech onsets

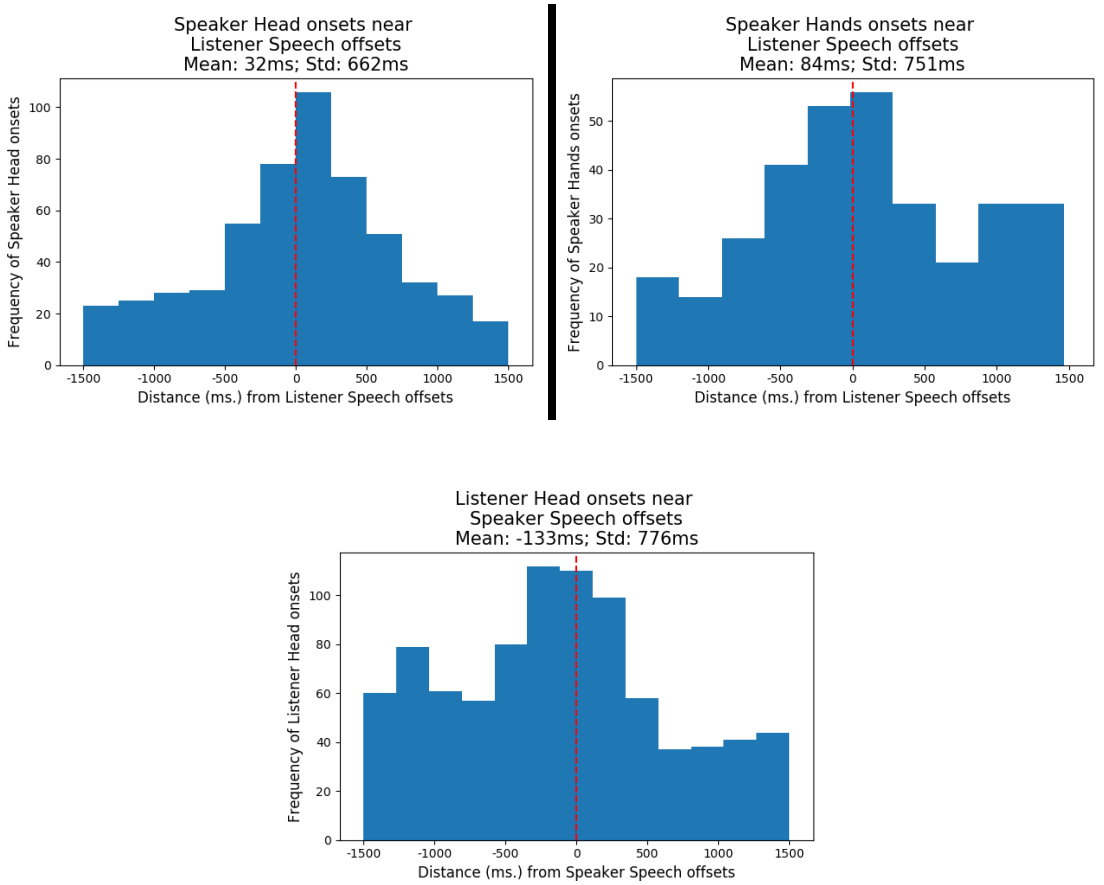


The other pairs of behavior onsets that exhibit some sort of timing relation are head and speech. For both listeners and speakers, head onsets are timed near speech onsets, although in the opposite directions. Speaker head onsets tend (slightly) to follow listener

speech onsets, while listener head onsets tend to precede listener speech onsets (although the rather uniform frequencies around these peaks indicate that there is a great deal more going on than just these timing relationships). For some listener back-channels and turns, speakers will immediately give a head response (usually a nod), which may account for the peak in Figure 62a. The peak in Figure 62b may depict the back-channel head nods that occur after a speaker speech segment, which is then followed by another speaker speech segment.

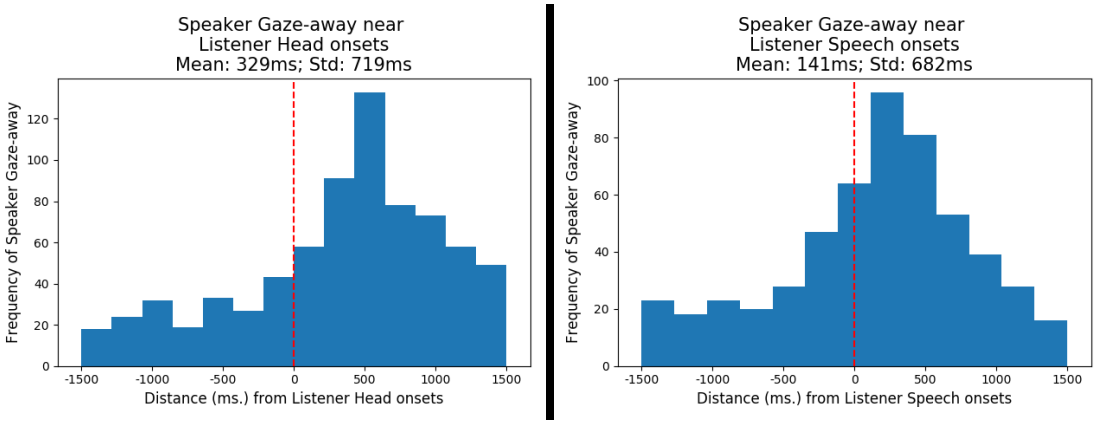
We now look at the onsets of behaviors near the offsets of other behaviors. A majority of these boundary pairs do not display any clear timing patterns. For example, listener gaze-towards does not pattern with any speaker offsets, listener and speaker speech onsets do not pattern clearly with speaker and listener manual gesture onsets. Others pattern together indirectly, or at least they do not pattern together as closely as the onsets of the two behaviors pattern together (e.g. listener head offsets follow speaker gaze-towards, but the listener head onsets seem to be the targeted boundary for this modality interaction).

Figure 63. Window histograms – Speaker head and manual gesture onsets near Listener speech offsets, and Listener head onsets near Speaker speech offsets



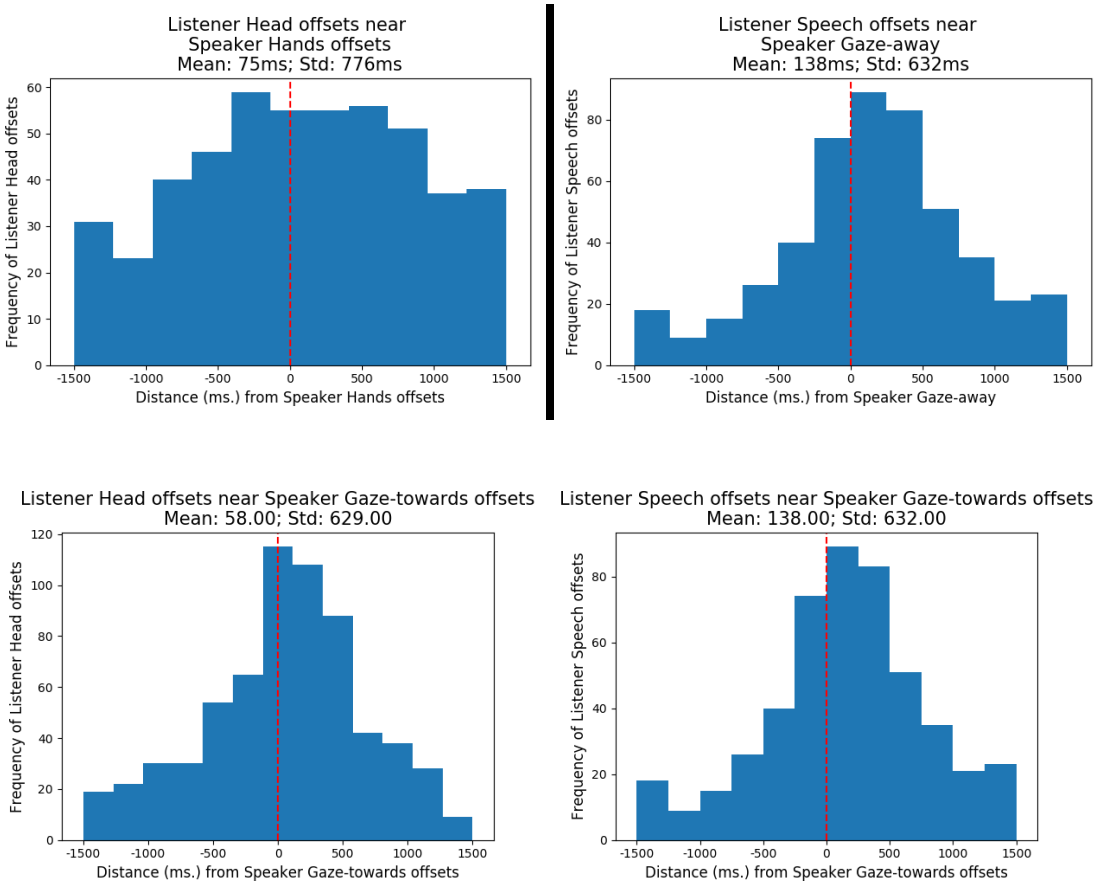
Speaker head and manual gesture are timed to co-occur near the offsets of listener speech, and listener head onsets are often timed to co-occur near the end of speaker speech. However, as we saw in Chapter 6 (Sections 2.3.2, 2.3.3), speech onsets are generally timed to follow speech offsets, and these head and manual gestures may simply be follows from their co-speech segments (most also exhibit less precise timing, although speaker head onsets near listener speech offsets are very nearly as precise as speaker speech onsets near listener speech offsets).

Figure 64. Window histogram – Speaker gaze-away near Listener head and speech onsets



Another interesting pattern is the timing relationship between listener speech and head gesture onsets near speaker gazes-away. The speaker gaze-shifts show a sharp peak in frequency just after the listener onsets (listener head gesture still preceding listener speech by a little). This looks like speakers are returning their gaze after checking for a response from the listener, but it may be more. Figures 64a/b below show the listener speech and head gesture offsets near speaker gaze-*away*, which show an every closer timing.

Figure 65. Window histogram – Listener head and speech offsets near Speaker gaze-towards



Listener head gesture offsets in particular are well-timed to match or follow speaker gaze-away. This suggests that speaker gaze-away is functioning as a cue, effectively communicating that the listener back-channel has been received, and the listener can then stop producing the back-channel. Signaling information to the interlocutor that lets them know their 'message' has been attended to, on a separate channel from the speech, seems functionally very like back-channeling, although since it is coming from the turn-taker, we might call it *front-channeling*.

3. Head + Speech

We now turn to the timing relations between subtypes of head and speech behaviors. As in previous chapters, this section will require more analysis than most, as head and speech have the most detailed coding schemes in this dataset. However, given that a number of these subtypes do not have a large number of tokens this also means that several subtypes do not have sufficient tokens to analyze in detail (listeners: single wags, multiple wags, multiple juts, multiple retractions; speakers: multiple juts, multiple retractions).

3.1 Likelihood Measures

3.1.1 Speaker Speech Subtypes and Listener Head Subtypes

We begin by looking at the likelihood measures of overlapping speaker speech subtypes and listener head subtypes. Table 52 below shows the conditional probabilities and odds ratios of the overlap between each listener head subtype and any speaker speech (speech turns and back-channels).

Table 52. Conditional probabilities and odds ratios of Listener head types with all Speaker speech

Listener head type	C. Prob. / S. Speech	C. Prob. / L. Head	L. Head odds-ratio
Tilt towards	0.006	0.792	**1.347*****
Multiple nod	0.112	0.769	**1.200*****
Nod up	0.006	0.671	0.719***
Tilt away + return	0.002	0.646	0.643***
Single retraction	0.002	0.630	0.601***
Single nod	0.021	0.622	0.573***
Tilt away	0.004	0.613	0.557***
Retraction back	0.003	0.604	0.536***
Nod down	0.006	0.578	0.480***
Jut in	0.002	0.562	0.453***
Tilt towards + return	0.002	0.491	0.339***
Single jut	0.002	0.474	0.317***
Single shake	0.002	0.450	0.288***
Multiple shake	0.005	0.399	0.231***

Looking at the conditional probabilities, we see that speaker speech overlaps with listener head gestures more than 50% of the time for most gestures, the exceptions being listener shakes and two single-cycle towards-and-return gestures. Shakes are the most likely co-turn head gesture for listeners (Chapter 5, Section 3.2) and single shakes are the most likely co-back-channel head gesture. Few of these gestures overlap with even 1% of speaker head gestures – even multiple nods only account for about 11%.

Despite the large degree of overlap for listener head gestures, only tilts towards and multiple nods are more likely than expected, with all other gesture being less likely. The multiple nod is a frequent back-channel, indicating comprehension or signaling attention.

The tilt towards is a repositioning gesture, thought to signify a shift in perspective, and a shift in perspective is an effective way to signal that one is paying attention to the message.

Table 53 looks at the overlap between listener head subtypes and the two most common forms of speaker speech turn, declaratives and interrogatives.

Table 53. Conditional probabilities and odds ratios of Listener heads with Speaker declarative and interrogative speech

Speaker declarative + Listener Head subtypes				**Speaker interrogative + Listener Head subtypes**			
Listener head type	**CP | L.**	**CP | S.**	**Odds-ratio**	**Listener head type**	**CP | L.**	**CP | S.**	**Odds-ratio**
Tilt towards	0.652	0.007	**1.782*****	**Multiple nod**	0.157	0.158	**1.671*****
Tilt away + return	0.607	0.003	**1.466*****	**Retraction back**	0.155	0.006	**1.541*****
Multiple nod	0.556	0.116	**1.211*****	**Jut in**	0.130	0.002	ns
Nod up	0.495	0.006	ns	**Single retraction**	0.126	0.002	ns
Single retraction	0.473	0.002	ns	**Single jut**	0.123	0.003	ns
Single nod	0.441	0.022	0.742***	**Single nod**	0.099	0.024	ns
Tilt away	0.437	0.005	0.736***	**Tilt away**	0.095	0.005	ns
Nod down	0.410	0.006	0.657***	**Nod up**	0.083	0.005	0.757**
Tilt towards + return	0.408	0.002	0.653***	**Nod down**	0.076	0.005	0.691***
Retraction back	0.373	0.003	0.564***	**Tilt towards + return**	0.044	0.001	0.391***
Multiple shake	0.361	0.006	0.532***	**Tilt towards**	0.030	0.002	0.255***
Single shake	0.360	0.002	0.532***	**Single shake**	0.022	0.001	0.190***
Jut in	0.325	0.001	0.455***	**Multiple shake**	0.007	0.001	0.056***
Single jut	0.276	0.001	0.361***	**Tilt away + return**	0.000	0.000	0.000***

Comparing speaker declaratives to speaker speech overall, we see that, in addition to tilts towards and multiple nods, tilts away+return are also more likely than expected. These are similar to tilts toward in that both finish with motion towards the interlocutor. While

single and multiple shakes remain unlikely, the least likely listener head gestures to co-occur with speaker declaratives are juts.

Juts are relatively more likely during speaker interrogatives. Here, multiple nods are even more likely that during declaratives (1.67 compared to 1.21), and retractions back are also more likely than expected. Indeed, all retractions and juts are among the most likely during interrogatives, and tilts and shakes are among the least likely.

Table 54 shows overlap information between listener heads and two less common speaker speech behaviors, fillers and incompletes.

Table 54. Conditional probabilities and odds ratios of Listener heads with Speaker filler and incomplete speech

Speaker filler + Listener Head subtypes				**Speaker incomplete + Listener Head subtypes**			
Listener head type	**CP | L.**	**CP | S.**	**Odds-ratio**	**Listener head type**	**CP | L.**	**CP | S.**	**Odds-ratio**
Jut in	0.061	0.005	**2.345*****	**Tilt towards**	0.079	0.005	ns
Tilt away + return	0.039	0.004	ns	**Single jut**	0.075	0.002	ns
Nod up	0.038	0.009	**1.419***	**Nod down**	0.064	0.005	0.678***
Tilt towards	0.032	0.006	ns	**Retraction back**	0.062	0.003	0.654**
Nod down	0.028	0.007	ns	**Single nod**	0.060	0.017	0.630***
Tilt away	0.024	0.005	ns	**Tilt away**	0.057	0.003	0.600***
Single nod	0.022	0.020	0.797*	**Nod up**	0.056	0.004	0.588***
Single shake	0.018	0.002	ns	**Single shake**	0.051	0.001	0.534**
Multiple nod	0.011	0.044	0.373***	**Jut in**	0.046	0.001	0.486**
Multiple shake	0.008	0.003	0.290***	**Multiple nod**	0.045	0.054	0.446***
Retraction back	0.001	0.000	0.051***	**Tilt towards + return**	0.038	0.001	0.398***
Single retraction	0.000	0.000	0.000***	**Single retraction**	0.032	0.001	0.325***
Tilt towards + return	0.000	0.000	0.000***	**Multiple shake**	0.023	0.002	0.238***
Single jut	0.000	0.000	0.000***	**Tilt away + return**	0.000	0.000	0.000***

As we can see from the conditional probabilities, the overlaps here are much less common than with declaratives and interrogatives, so we should be cautious in extrapolating from this data. During fillers, almost all listener gesture is less likely than expected, but we do see that repositioning gestures such as juts in and nods up are more likely than expected. This is interesting given that, as we saw in Chapter 5, Section 3.1.1, repositioning gestures are also the most common speaker head gestures during speaker fillers. Listener may be imitating speakers, or they may also be shifting their perspective as the narrative shifts.

During speaker incompletes, there are no listener head gestures that are more likely than expected – the only gestures that are even at chance likelihood are repositioning gestures with motion-towards: tilts towards and juts in.

Overall, there is an interesting similarity between these across-role co-occurrence patterns and the within-role patterns we saw in Chapter 5. That is, co-speech head gesture for speakers is more common during declaratives and interrogatives, and less common during fillers and incompletes. Here, we see that listener head gesture is also more common during speaker declaratives and interrogatives, and less common during fillers and incompletes. This may not be perfectly clear from the above tables, because they focus on head subtypes, but if consider the distributions of tokens, listener multiple nods are by far the most frequent listener head gesture, and these are only more likely than expected during declaratives and interrogatives.

If these listener head gestures are responding to speaker cues, it is still not clear whether they are responding more to the increased head gestures or to the nature of the message in the speech. This could be a fruitful place to do more qualitative analysis.

Table 55. Conditional probabilities and odds ratios of Listener heads with Speaker back-channels

Listener head type	C. Prob. / L. Head	C. Prob. / S. Back-channel	Back-channel odds-ratio
Tilt towards + return	0.095	0.029	**11.433*****
Retraction back	0.050	0.022	**5.693*****
Single shake	0.042	0.012	**4.687*****
Single jut	0.019	0.006	**2.104***
Multiple shake	0.019	0.017	**2.043*****
Single nod	0.014	0.037	**1.484****
Multiple nod	0.012	0.137	**1.321*****
Nod down	0.012	0.009	ns
Single retraction	0.009	0.002	ns
Tilt away	0.008	0.004	ns
Nod up	0.007	0.005	ns
Jut in	0.000	0.000	ns
Tilt towards	0.000	0.000	0.000***
Tilt away + return	0.000	0.000	0.000*

During speaker back-channels, listener head gestures are much more common, presumably both because listeners are taking the role of speaker (and with it the co-speech head gesture), and because they are responding to the back-channels. Both away-motion head gestures (tilts toward+return and retraction back) are common here, as are both kinds of shakes, and multiple and single nods. And, where shakes are more likely here, tilts toward and tilts away+return (the most likely during declaratives) are the least likely here, while most repositioning gestures are at chance.

Although there are too few tokens of many speaker back-channels, there are sufficient speaker laughs and affirmations to examine in detail (Table 56).

Table 56. Conditional probabilities and odds ratios of Listener heads with Speaker laughs and affirmations

Speaker laugh + Listener Head subtypes				**Speaker affirmation + Listener Head subtypes**			
Listener head type	**CP | L. Head**	**CP | S. Laugh**	**Odds-ratio**	**Listener head type**	**CP | L. Head**	**CP | S. Aff.**	**Odds-ratio**
Multiple shake	0.047	0.120	**16.19*****	**Retraction back**	0.031	0.020	**5.005*****
Jut in	0.026	0.015	**7.824*****	**Tilt towards + return**	0.022	0.010	**3.469*****
Nod down	0.014	0.029	**4.199*****	**Single jut**	0.019	0.008	**3.015****
Single shake	0.013	0.010	**3.898****	**Nod down**	0.012	0.012	**1.784***
Single nod	0.011	0.083	**3.509*****	**Multiple nod**	0.010	0.157	**1.560*****
Tilt towards	0.000	0.000	ns	**Multiple shake**	0.009	0.012	ns
Tilt away	0.000	0.000	ns	**Single retraction**	0.009	0.003	ns
Tilt towards + return	0.000	0.000	ns	**Tilt away**	0.008	0.006	ns
Tilt away + return	0.000	0.000	ns	**Nod up**	0.007	0.007	ns
Single jut	0.000	0.000	ns	**Single nod**	0.005	0.018	ns
Retraction back	0.000	0.000	ns	**Single shake**	0.000	0.000	ns
Single retraction	0.000	0.000	ns	**Tilt away + return**	0.000	0.000	ns
Multiple nod	0.001	0.029	0.245***	**Jut in**	0.000	0.000	ns
Nod up	0.000	0.000	0.000*	**Tilt towards**	0.000	0.000	0.000**

Speaker affirmations make up the bulk of speaker back-channels, so their results are not strikingly different from Table 55, except that only single nods are more likely, while multiple nods are much less likely than expected, and the retraction back and tilt towards+return are no longer more likely than expected. These last two are the most likely to co-occur with speaker laughs, however, along with single and multiple nods, and single juts.

3.1.2 Listener Speech Subtypes and Speaker Head Subtypes

Having examined the co-occurrence patterns of speaker speech and listener heads, we will now examine the reverse. Because listener turns are less frequent and listener back-

channels are more frequent, these charts will be skewed towards back-channels rather than turns. We begin with likelihood measures of the overlaps between different speaker head subtypes and all listener speech (Table 57).

Table 57. Conditional probabilities and odds ratios of Speaker heads with all Listener speech

Speaker head types	C. Prob. / L. Speech	C. Prob. / S. Head	L. Speech odds-ratio
Retraction back	0.013	0.106	**1.417*** **
Multiple wag	0.012	0.098	**1.284****
Multiple nod	0.075	0.088	**1.157*****
Multiple shake	0.070	0.087	**1.144*****
Single nod	0.045	0.076	ns.
Nod up	0.020	0.087	ns.
Tilt away	0.015	0.078	ns.
Tilt towards + return	0.009	0.071	ns.
Tilt away + return	0.006	0.081	ns.
Single retraction	0.008	0.069	ns.
Single shake	0.016	0.068	0.862*
Nod down	0.022	0.067	0.850**
Single jut	0.011	0.062	0.784**
Tilt towards	0.011	0.053	0.657***
Jut in	0.006	0.037	0.453***
Single wag	0.001	0.016	0.198***

Of the few speaker head behaviors that are more likely than expected to co-occur with listener speech, none are especially likely, although all multiple cycles of gestures are slightly more likely than expected, and retractions back are more likely. But most head gestures are at chance, or not far from being at chance, the exceptions being tilts towards, juts in, and single wags.

Looking in more depth at speaker heads co-occurring with listener speech, we will start with listener turns. Table 58 shows these overlap relations for listener declaratives and interrogatives.

Table 58. Conditional probabilities and odds ratios of Speaker heads with Listener declarative and interrogative speech

Listener declarative + Speaker Head subtypes				**Listener interrogative + Speaker Head subtypes**			
Speaker head type	**CP | S.**	**CP | L.**	**Odds-ratio**	**Speaker head type**	**CP | S.**	**CP | L.**	**Odds-ratio**
Multiple nod	0.084	0.157	**2.765*** **	**Multiple nod**	0.030	0.096	**1.504*** **
Tilt towards	0.070	0.033	**2.055*** **	**Single nod**	0.025	0.055	**1.215***
Multiple shake	0.040	0.070	**1.125***	**Tilt towards**	0.021	0.017	ns.
Tilt away	0.039	0.016	ns.	**Single retraction**	0.017	0.008	ns.
Retraction back	0.034	0.009	ns.	**Tilt towards + return**	0.016	0.007	ns.
Tilt away + return	0.033	0.006	ns.	**Nod down**	0.014	0.017	0.661***
Tilt towards + return	0.029	0.008	ns.	**Tilt away**	0.010	0.007	0.492***
Multiple wag	0.029	0.008	ns.	**Multiple shake**	0.008	0.023	0.342***
Single shake	0.026	0.014	0.728**	**Single shake**	0.004	0.004	0.203***
Single retraction	0.026	0.007	0.726*	**Jut in**	0.001	0.001	0.063***
Nod down	0.026	0.018	0.704***	**Nod up**	0.001	0.001	0.045***
Single nod	0.025	0.032	0.683***	**Single jut**	0.001	0.001	0.039***
Nod up	0.025	0.012	0.684***	**Retraction back**	0.000	0.000	0.000***
Single jut	0.015	0.006	0.395***	**Tilt away + return**	0.000	0.000	0.000***
Jut in	0.014	0.005	0.369***	**Multiple wag**	0.000	0.000	0.000***
Single wag	0.000	0.000	0.000***	**Single wag**	0.000	0.000	0.000***

For listener declaratives, most of these overlaps make up only a small proportion of either behavior. Even the greatest proportions are less than 20%, and mostly less than 10%. During listener declaratives, speakers tend to use multiple nods, multiple shakes, and tilts towards, the latter of which seems to indicate a shift in consideration from the speaker's message to the listener's. Other tilts are mostly at chance, along with retractions back. Quite uncommon are juts (and single retractions), other nods, and single wags.

During listener interrogatives, only speaker single and multiple nods are more likely than expected. When listeners ask a question, speakers very often begin nodding in response well before the question is completed. Wags and away-motion gestures never co-occur here.

There are far more listener back-channels than listener speech turns, so we can look at the co-occurrence patterns for speaker heads with most types of back-channel (although there is too little interaction with collaborative finishes and newsmarkers to examine). However, interpreting these co-occurrence patterns is somewhat problematic, as it is easier to interpret listener back-channels as responding to speech content, rather than head content, and so patterns seen here may be due to how speaker head gestures co-vary with different kinds of speech content. Still, these patterns may be helpful in future work to identify such patterns of speech and head covariance. We look first at speaker head subtypes and listener back-channels (Table 59).

Table 59. Conditional probabilities and odds ratios of Speaker heads with Listener back-channels

Listener back-channel + Speaker Head subtypes			
Speaker head type	**CP | S. Head**	**CP | L. Back-channel**	**Odds-ratio**
Retraction back	0.073	0.011	ns.
Nod up	0.070	0.020	**1.155***
Multiple nod	0.068	0.074	**1.125***
Multiple shake	0.066	0.067	ns.
Single nod	0.062	0.047	ns.
Single shake	0.061	0.019	ns.
Single retraction	0.059	0.009	ns.
Multiple wag	0.056	0.009	ns.
Tilt towards + return	0.051	0.008	ns.
Single jut	0.049	0.011	0.787***
Tilt away	0.047	0.011	0.750**
Tilt towards	0.046	0.012	0.732***
Tilt away + return	0.039	0.004	0.625**
Nod down	0.039	0.016	0.613***
Jut in	0.032	0.007	0.505*
Single wag	0.016	0.001	0.256***

Looking at back-channels as a whole, we see very little in the way of attracting speaker heads. Speaker nods up and multiple nods are significantly more likely, but are only barely greater than chance. In the opposite direction, however, we see several towards-motion gestures that are less likely than expected, as well as single wags.

In Table 60, we look at the overlaps of speaker heads and two of the most common listener back-channels: acknowledgments and assessments.

Table 60. Conditional probabilities and odds ratios of Speaker heads with Listener acknowledgments and assessments

Listener acknowledgment + Speaker Head				**Listener assessment + Speaker Head**			
Speaker head type	**CP | S. Head**	**CP | L. Ack.**	**Odds-ratio**	**Speaker head type**	**CP | S. Head**	**CP | L. Ass.**	**Odds-ratio**
Multiple wag	0.036	0.022	**2.36*****	**Retraction back**	0.050	0.015	**1.641*****
Single retraction	0.034	0.020	**2.20*****	**Multiple nod**	0.034	0.072	ns
Tilt towards	0.025	0.026	**1.61*****	**Nod up**	0.031	0.018	ns
Tilt away	0.024	0.023	**1.58****	**Tilt towards + return**	0.031	0.009	ns
Single shake	0.024	0.029	**1.55*****	**Multiple shake**	0.029	0.057	ns
Multiple nod	0.021	0.089	**1.38*****	**Single nod**	0.028	0.041	ns
Single nod	0.020	0.059	**1.31****	**Multiple wag**	0.020	0.006	0.62***
Nod up	0.020	0.023	Ns	**Jut in**	0.019	0.008	0.58***
Single jut	0.020	0.018	Ns	**Nod down**	0.018	0.015	0.57***
Multiple shake	0.017	0.066	Ns	**Tilt towards**	0.018	0.010	0.57***
Single wag	0.015	0.005	Ns	**Tilt away**	0.018	0.008	0.56***
Retraction back	0.014	0.009	Ns	**Single shake**	0.017	0.010	0.54***
Tilt away + return	0.012	0.005	Ns	**Tilt away + return**	0.015	0.003	0.48**
Nod down	0.011	0.018	0.70*	**Single retraction**	0.013	0.004	0.42***
Jut in	0.006	0.005	0.40***	**Single jut**	0.005	0.002	0.17***
Tilt towards + return	0.002	0.001	0.12***	**Single wag**	0.001	0.000	0.04***

Listener acknowledgments are more likely than expected to overlap with a number of speaker head gestures, although it is difficult to identify any shared features. Single shakes are common but multiple shakes are not; multiple wags are common but single wags are not; three half-cycles are more likely, but the other three are not. Speaker head gestures during listener acknowledgments are easier to interpret. Here, there are no head gestures that are more likely to overlap than expected, except the retraction back. The retraction back is often co-produced during dramatic narrative events, and assessments are a natural response to this.

Table 61. Conditional probabilities and odds ratios of Speaker heads with Listener affirmations

Listener affirmation + Speaker Head			
	CP | S. Head	**CP | L. Aff.**	**Odds-ratio**
Single shake	0.010	0.046	**2.540*****
Single jut	0.007	0.025	**1.827***
Tilt towards + return	0.006	0.015	ns
Multiple shake	0.005	0.086	**1.413***
Multiple nod	0.005	0.089	**1.381***
Nod up	0.004	0.019	ns
Retraction back	0.004	0.010	ns
Jut in	0.004	0.013	ns
Tilt away	0.004	0.013	ns
Nod down	0.002	0.013	ns
Single wag	0.000	0.000	ns
Tilt away + return	0.000	0.000	0.000**
Single retraction	0.000	0.000	0.000**
Single nod	0.000	0.000	0.000***
Tilt towards	0.000	0.000	0.000***
Multiple wag	0.000	0.000	0.000**

During listener affirmations, interestingly, speakers are more likely than expected to be nodding or shaking their heads. This is interesting because we saw that listeners producing affirmations are also more likely than expected to be nodding or shaking. Given that affirmations are positive and shakes are often interpreted as negative, this is a combination we might not have expected to find.

Finally, in Table 62, we look at speaker head overlaps with listener laughs.

Table 62. Conditional probabilities and odds ratios of Speaker heads with Listener laughs

	CP / S. Head	CP / L. Laugh	Laugh odds-ratio
Tilt away + return	0.042	0.018	**3.108*****
Multiple wag	0.042	0.029	**3.111*****
Retraction back	0.034	0.023	**2.479*****
Tilt away	0.028	0.029	**2.070*****
Nod down	0.024	0.043	**1.761*****
Tilt towards + return	0.021	0.014	**1.476***
Nod up	0.019	0.024	**1.355***
Multiple shake	0.018	0.078	**1.275****
Single jut	0.013	0.013	ns
Single nod	0.013	0.042	ns
Multiple nod	0.012	0.057	ns
Single retraction	0.010	0.007	ns
Single shake	0.009	0.012	0.607**
Jut in	0.005	0.005	0.346***
Tilt towards	0.003	0.004	0.242***
Single wag	0.000	0.000	0.000***

Half of speaker head subtypes are more likely than expected to overlap with listener laughs, although there is not a clear pattern unifying the head behaviors that are more or less likely than expected.

3.2 N-grams

Having examined the patterns of overlap in Section 3.1, we now turn to sequential patterns of onsets and offsets (Table 63).

Table 63. Speaker and Listener Head and Speech boundaries (1-second window)

Roles	Type	Bigram (1 + 2)	Frequency	Symm. CP
L + **S**	off + **off**	L. Head offsets + **S. Speech offsets**	167	0.015
L + **S**	on + **off**	L. Head onsets + **S. Speech offsets**	155	0.013
S + L	**off** + on	**S. Head offsets** + L. Speech onsets	116	0.010
L + **S**	off + **on**	L. Speech offsets + **S. Head onsets**	110	0.009
S + L	**on** + off	**S. Speech onsets** + L. Head offsets	128	0.009
S + L	**off** + off	**S. Speech offsets** + L. Head offsets	119	0.008
L + **S**	off + **on**	L. Head offsets + **S. Speech onsets**	118	0.008
S + L	**off** + on	**S. Speech offsets** + L. Head onsets	111	0.007
S + L	**on** + on	**S. Speech onsets** + L. Head onsets	106	0.006
S + L	**on** + off	**S. Head onsets** + L. Speech offsets	86	0.006
L + **S**	off + **off**	L. Speech offsets + **S. Head offsets**	80	0.005
L + **S**	on + **on**	L. Head onsets + **S. Speech onsets**	89	0.004
L + **S**	on + **off**	L. Speech onsets + **S. Head offsets**	71	0.004
S + L	**off** + off	**S. Head offsets** + L. Speech offsets	64	0.003
L + **S**	on + **on**	L. Speech onsets + **S. Head onsets**	61	0.003
S + L	**on** + on	**S. Head onsets** + L. Speech onsets	59	0.003

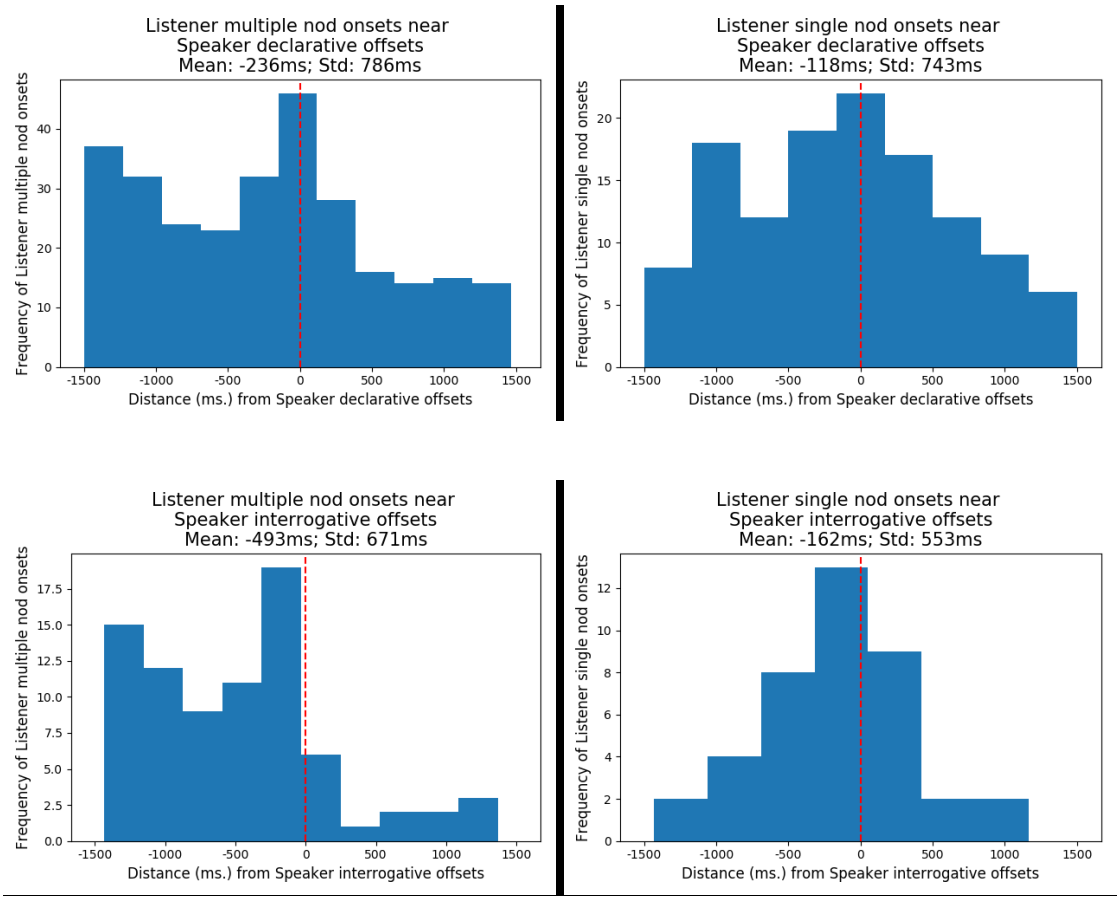
Within the 1-second window, we see that the most frequent bigram pairs involve listener head onsets and offsets being followed by speaker speech – these will be back-channel responses to the speaker's speech. Among the next most common are onsets that follow offsets, of both speech following head offsets and heads following speech offsets. The least common are onsets following onsets, although listener head onsets following speaker speech onsets are more common than the others.

3.3 Window Histograms

Looking at the frequency distributions of across-role head and speech boundaries near each other, we do not see a great many clear patterns. The one repeated pattern that we do

see is the timing of single and multiple nods relative to the interlocutor's speech offsets. We know that many nods that begin near interlocutor speech offsets are accompanying speech onsets, and we know that speech onsets follow speech offsets very precisely across roles (Chapter 5, Sections 2.3.2, 2.3.3). In Figures 66a-d below, we can see the temporal distribution of listener head onsets near speaker speech offsets.

Figure 66. Window histogram – Listener nod onsets near Speaker speech offsets



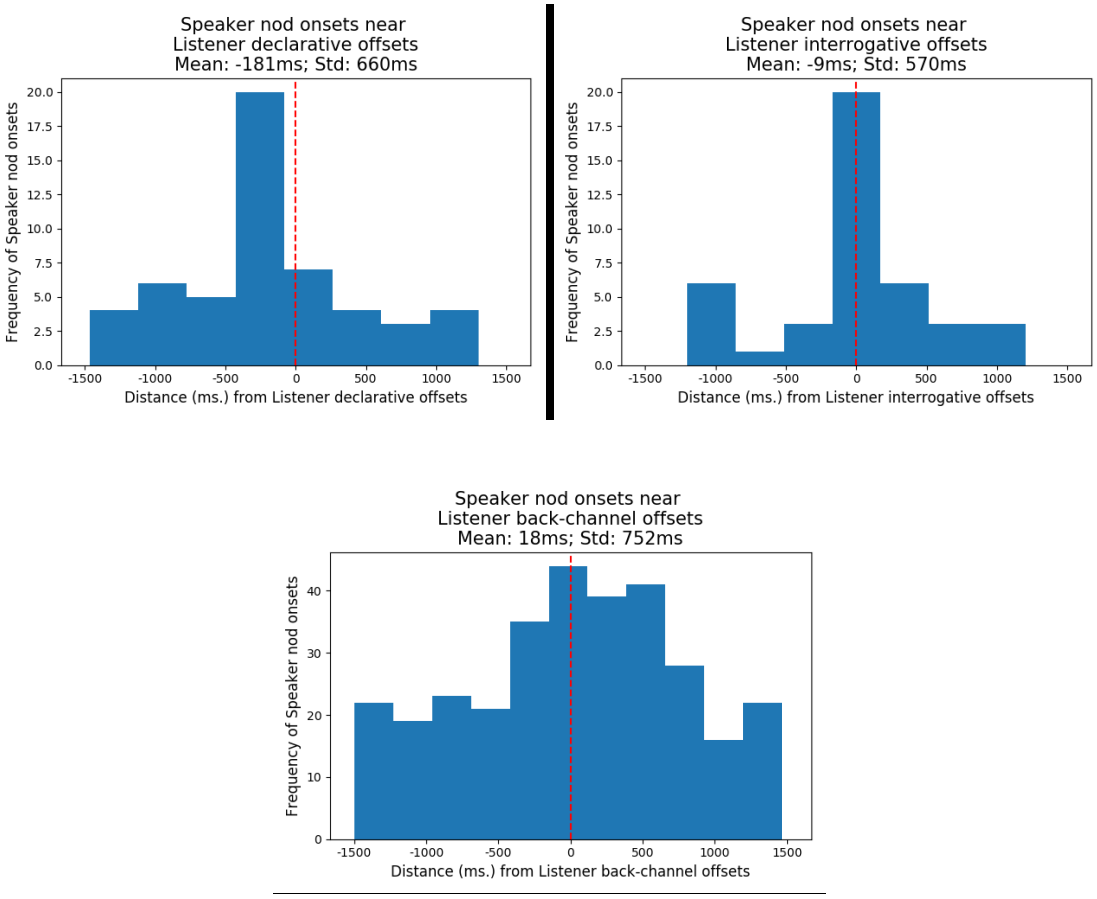
Listener nods following speaker declarative offsets do show a frequency peak at the moment of the offset, but it is not nearly as precise as the speech onsets.

Following speaker interrogatives, there is slightly more precision. Listener multiple nods can occur at various points before the offset of the interrogative, but they peak at the

speech's offset, and tend not to follow it. Single nods also show a clear peak just before the interrogative offset. Overall, there is a great deal less precision that between speech and speech, as also seen in the n-grams in Section 3.2.

Looking at speaker head nods near listener speech offsets, however, we see something a different pattern (Figures 67a-c).

Figure 67. Window histogram – Speaker nod onsets near Listener speech offsets



Speaker nod onsets near listener declarative and interrogative offsets are not only timed with great precision (tending to peak just before the offset for declaratives and just at or after for interrogatives), but they are rarely produced away from the speech offset. It's not entirely clear why there is such a dramatic difference across roles. It's possible that,

because listener turns are so relatively short and infrequent, speakers do not have time to produce all the back-channels nods that fill up the variance in Figures 66a-d. Or it may be that speakers are not interested in listeners taking long turns, and are simply waiting till they can affirm whatever the listener has said and continue on with their story.

However, in Figure 67c, where we see speaker nods occurring near listener back-channels, we see nothing of the same precision, but rather a gentler peak on top of a flat distribution of nods. Many listener back-channels are produced during speaker speech, which explains the flatness. However, the peak seems to be evidence that speakers are indeed responding to some listener back-channels.

4. Head+Gaze

In Section 2, we looked at the co-occurrence patterns of gaze and other modalities, where we saw that both speaker and listener head behaviors were more likely during interlocutor gaze-towards. We broke down gaze by its different combinations with interlocutor gaze, and found, for example, that speaker gaze is more likely than expected during mutual gaze, but less likely than expected when only the listener is gazing towards the speaker. In this section, we will look at the interaction between head and gaze across role, continuing to look at gaze vs. mutual gaze, but also breaking down head into its different subtypes.

4.1 Likelihood Measures

We begin by looking at likelihood measures. Table 64 shows the conditional probabilities and odds-ratios of the overlaps of listener head subtypes with speaker gaze, and also with mutual gaze.

Table 64. Conditional probabilities and odds-ratios of Speaker/Mutual gaze-towards with Listener heads

Listener Head + Speaker Gaze				**Listener Head + Mutual Gaze**			
	CP. / S. Gaze	**CP / L. Head**	**Odds-ratio**		**CP | Head**	**CP | Gaze**	**Odds-ratio**
Jut in	0.003	0.583	**1.986*****	**Single jut**	0.547	0.005	**3.243*****
Multiple nod	0.133	0.511	**1.557*****	**Multiple nod**	0.529	0.208	**3.526*****
Retraction back	0.005	0.492	**1.379****	**Retraction back**	0.461	0.007	**2.290*****
Single retraction	0.002	0.456	ns.	**Nod up**	0.416	0.010	**1.912*****
Tilt towards	0.005	0.406	ns.	**Tilt towards + return**	0.406	0.004	**1.830*****
Single nod	0.024	0.394	0.920*	**Single shake**	0.373	0.004	**1.592*****
Nod up	0.006	0.364	0.812*	**Single nod**	0.353	0.033	**1.475*****
Single jut	0.002	0.347	0.755**	**Jut in**	0.351	0.003	**1.444*****
Tilt away + return	0.002	0.343	0.740**	**Tilt away**	0.345	0.007	**1.407*****
Tilt away	0.004	0.280	0.552***	**Tilt towards**	0.311	0.006	**1.208***
Nod down	0.005	0.268	0.517***	**Multiple shake**	0.280	0.009	ns.
Single shake	0.002	0.238	0.444***	**Tilt away + return**	0.271	0.003	ns.
Multiple shake	0.004	0.189	0.328***	**Single retraction**	0.258	0.002	ns.
Tilt towards + return	0.001	0.172	0.294***	**Nod down**	0.117	0.003	0.350***

Looking first at listener head subtypes overlapping with speaker gaze, we see that only a few types of behavior are more likely than expected (juts in, multiple nods, and retractions back), while most head gestures are less likely than expected (especially shakes and away-motion tilts). Multiple nods are the most frequent listener nod, and are often quite lengthy, so it says something that these are still more likely than expected, and suggests they are not being produced only as a symptom of comprehension.

We compare this to mutual gaze, where most listener head gestures are more likely than expected, and only nods down are less likely. This suggests that listeners are not only

concerned with their head gestures being seen, but also with seeing that they have been seen.

One interesting point to note about this nod down is that it is the only nod that is less likely than expected during both listener gaze-towards and mutual gaze. Back in Chapter 1 we discussed the many possible functions of the nod, two critical functions being 1) a signal of attention, and 2) a symptom of comprehension. Essentially, when is a nod communicative, and when is it merely informative? One good test of communicativeness is that a non-verbal cue tends to be used when it can be detected (and even better when one can see that it has been detected. Of all the four types of nods in this dataset, the nod down is the only type of nod that is not only not more likely during mutual gaze, but is quite a lot less likely than expected. This is not conclusive, but is suggestive that this form of nod is less often produced to be communicative, and possibly more often is a symptom of comprehension.

In Table 65, we look at speaker head subtypes overlapping with listener gaze and with mutual gaze.

Table 65. Conditional probabilities and odds-ratios of Listener/Mutual gaze-towards with Speaker heads

Speaker Head + Listener Gaze				Speaker Head + Mutual Gaze			
	CP | L. Gaze	**CP / S. Head**	**Odds-ratio**		**CP | Head**	**CP | Gaze**	**Odds-ratio**
Multiple wag	0.011	0.972	**6.184*****	**Jut in**	0.499	0.024	**2.702*****
Retraction back	0.011	0.955	**3.763*****	**Multiple wag**	0.437	0.016	**2.094*****
Single jut	0.016	0.949	**3.333*****	**Multiple nod**	0.435	0.106	**2.178*****
Multiple shake	0.069	0.928	**2.377*****	**Single retraction**	0.400	0.014	**1.792*****
Tilt away + return	0.007	0.920	**2.033*****	**Retraction back**	0.396	0.014	**1.762*****
Jut in	0.014	0.919	**2.039*****	**Multiple shake**	0.392	0.090	**1.791*****
Single shake	0.020	0.909	**1.789****	**Tilt away + return**	0.373	0.008	**1.591*****
Single wag	0.005	0.908	**1.759****	**Single jut**	0.366	0.019	**1.552*****
Single retraction	0.010	0.889	**1.425****	**Single nod**	0.343	0.058	**1.421*****
Tilt towards	0.017	0.856	ns.	**Tilt towards + return**	0.321	0.011	**1.265****
Single nod	0.046	0.856	ns.	**Tilt towards**	0.314	0.019	**1.227****
Multiple nod	0.066	0.843	ns.	**Single shake**	0.309	0.021	**1.201***
Tilt towards + return	0.009	0.827	0.849*	**Single wag**	0.300	0.005	ns.
Nod down	0.024	0.824	0.824**	**Tilt away**	0.271	0.014	ns.
Nod up	0.017	0.816	0.785**	**Nod down**	0.252	0.023	0.900*
Tilt away	0.014	0.816	0.781**	**Nod up**	0.194	0.013	0.637*

In contrast to listener heads, many speaker head subtypes are more likely than expected during listener gaze, although two of the most frequent types, single and multiple nods, are at chance. During mutual gaze, however, these two gestures are more likely than expected, multiple nods being more than twice as likely (2.18). The gestures that are less likely than expected are mostly half-cycle repositioning gestures.

4.2 N-grams

We look now at the sequential patterns of gaze and head gesture across roles. Table 66 shows the frequencies of the bigram pairs of gaze and head boundaries for both roles.

Table 66. Speaker and Listener Head and Gaze boundary bigrams (1-second window)

Bigram (1 + 2)	Frequency	Symm. CP	CP: 1|2	CP: 2|1
S. Gaze-away + L. Head offsets	202	0.043	0.194	0.220
S. Gaze-towards + L. Head onsets	190	0.037	0.183	0.203
L. Head onsets + **S. Gaze-away**	123	0.016	0.134	0.119
L. Head offsets + **S. Gaze-away**	112	0.013	0.122	0.107
S. Head offsets + L. Gaze-away	84	0.008	0.200	0.042
L. Gaze-towards + **S. Head onsets**	82	0.008	0.041	0.191
S. Head onsets + L. Gaze-towards	75	0.007	0.175	0.038
L. Gaze-away + **S. Head onsets**	69	0.006	0.035	0.164
L. Gaze-towards + **S. Head offsets**	69	0.006	0.035	0.161
L. Gaze-away + **S. Head offsets**	63	0.005	0.032	0.150
L. Head offsets + **S. Gaze-towards**	65	0.004	0.069	0.062
S. Gaze-away + L. Head onsets	63	0.004	0.061	0.069
L. Head onsets + **S. Gaze-towards**	60	0.004	0.064	0.058
S. Head offsets + L. Gaze-towards	54	0.003	0.126	0.027
S. Head onsets + L. Gaze-away	40	0.002	0.095	0.020
S. Gaze-towards + L. Head offsets	35	0.001	0.034	0.037

The bigrams that show the greatest dependency (which with these binary categories amounts to frequency) within this window involve listener head offsets near speaker gaze-away, or listener head onsets near speaker gaze-towards. For both of these kinds of pairs, it is the speaker gaze boundary that is more likely to precede the listener head boundary. So when speakers gaze away, listener complete their head gestures, and when speakers gaze towards, listener begin them. Listeners do also complete their head

gestures prior to listeners gazing away, but at just over half the rate of the reverse (112 to 202), and listener head onsets do precede speaker gaze-towards, but at less than a third the rate of the reverse (60 to 190). Another frequent bigram involves speakers beginning a head gesture, then speakers gazing away (123).

These findings suggest that the dominant pattern is for speakers to gaze towards the listener, the listener to begin a head gesture, and then finish it after (or before) the speaker gazes away. But to see these more clearly, we will look at 4-grams of these behavior boundaries (Table 67). Only the most frequent 4-grams are displayed here, and only 4-grams that involve the beginning and completion of a gaze-shift and head gesture for speaker and listener.

Table 67. Speaker and Listener Head and Gaze boundary 4-grams (3-second window)

4-grams	Frequency
S. Gaze-towards + L. Head onsets + **S. Gaze-away** + L. Head offsets	40
L. Gaze-away + **S. Head onsets** + L. Gaze-towards + **S. Head offsets**	23
L. Head onsets + **S. Gaze-away** + L. Head offsets + **S. Gaze-towards**	18
S. Gaze-towards + L. Head onsets + L. Head offsets + **S. Gaze-away**	17
S. Gaze-away + L. Head offsets + **S. Gaze-towards** + L. Head onsets	10
S. Head onsets + L. Gaze-away + **S. Head offsets** + L. Gaze-towards	9
S. Head onsets + L. Gaze-towards + **S. Head offsets** + L. Gaze-away	9
L. Head offsets + **S. Gaze-towards** + L. Head onsets + **S. Gaze-away**	9
L. Gaze-away + **S. Head offsets** + L. Gaze-towards + **S. Head onsets**	8
S. Head offsets + L. Gaze-away + **S. Head onsets** + L. Gaze-towards	8
L. Head onsets + **S. Gaze-towards** + **S. Gaze-away** + L. Head offsets	8

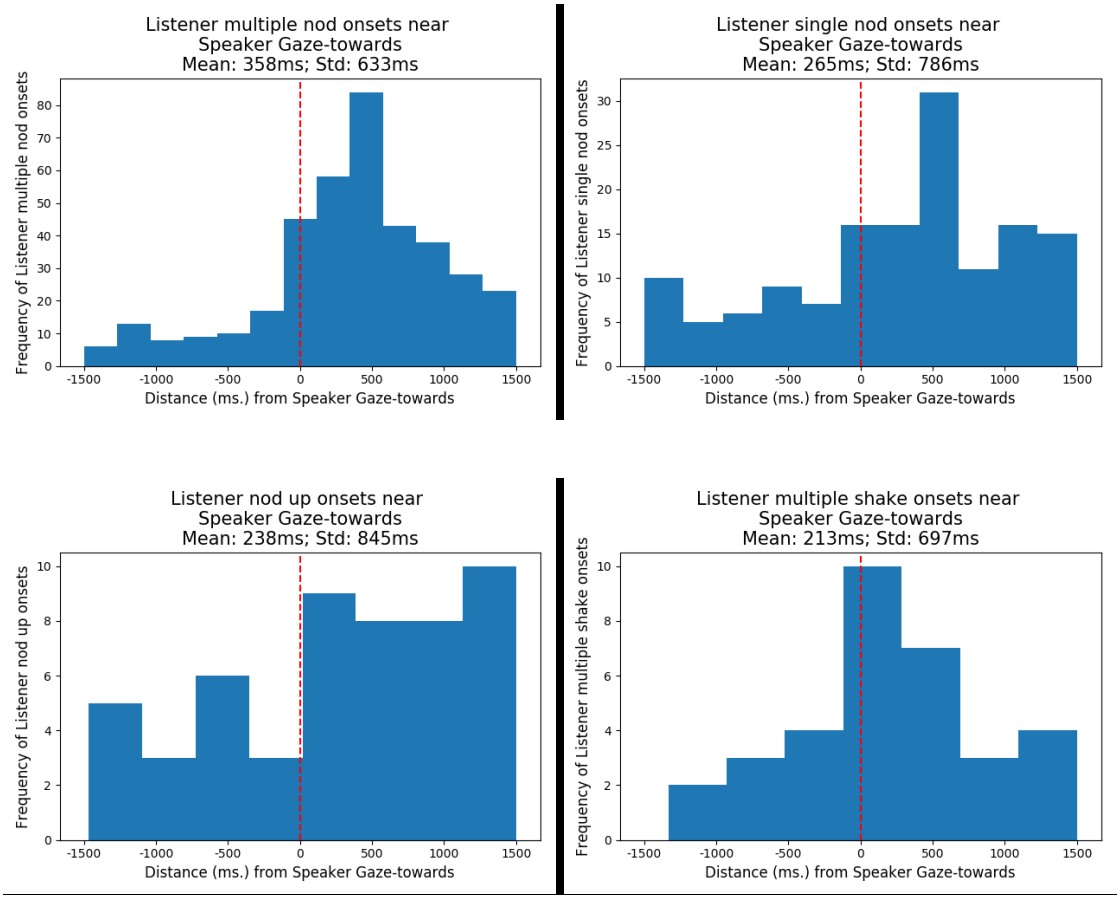
As we can see, the most frequent pattern is precisely what was predicted. The second most frequent pattern is somewhat the reverse, in which listeners gaze away, followed by

the onset of a speaker head gesture, followed by the listener returning their gaze, and the speaker finishing their head gesture. These are primarily listener turns, where the listener looks away as they begin to speak, the speaker nods an affirmation, and the listener looks back before the nod is complete.

4.3 Window Histograms

In looking at the frequency distributions of head and gaze boundaries across roles, there are not a lot of clear timing patterns, because speakers do not time many of their head gestures relative to listener gaze, and, while listeners do time head gestures to listener gaze-towards, these are primarily multiple nods. Listener head gesture onsets near speaker gaze-towards are shown below in Figures 68a-d.

Figure 68. Window histogram – Listener nod onsets near Speaker gaze-towards



As we can see, listener multiple nods are timed to follow the onset of speaker gaze towards. Single nods seem to follow the same pattern, and, to a less clear extent, nods up and multiple shakes. While these heads and gaze behaviors show a similar pattern to the pattern seen between speaker speech offsets and listener speech onsets, the timing differs. For the speech onsets, the frequency rises rapidly at the speech offset, and then rapidly diminishes. With gaze-shift and head onsets, however, there is a less sharp incline (but still a clear one, at least for multiple nods), but the decline is not nearly as precipitous. Listeners are more apt to take their time (around 700 to 1000ms) with their heads than they are with their speech (around 200ms). This may be because there are better cues in the speech stream to let listeners know when to time their verbal response than there are

in the non-verbal. Or it may be that listeners know they have more time during a speaker glance than they do during a speaker pause.

5. Gaze+Speech

We saw in Section 2.1 that both speaker and listener gaze are more likely than expected to overlap with interlocutor gaze. And, unlike with speaker head gesture, this was true regardless of whether there was mutual gaze or only listener gaze-towards. In this section we will continue to look at interlocutor gaze compared with mutual gaze, but will also break down speech into its different subtypes.

5.1 Likelihood Measures

Tables 7.18 and 7.19 show the conditional probabilities and odds-ratios of the overlaps between listener speech types (speech turns and back-channels) and speaker gaze (mutual and non-mutual).

Table 68. Conditional probabilities and odds-ratios of Speaker/Mutual Gaze with Listener Speech-types

Listener Speech turn-types + Speaker Gaze				**Listener Speech turn-types + Mutual Gaze**			
	CP | Gaze	**CP | Speech**	**Odd-ratio**		**CP | Gaze**	**CP | Speech**	**Odds-ratio**
Incomplete	0.007	0.725	**5.673*****	**Interrogative**	0.033	0.429	**2.044*****
Declarative	0.070	0.627	**3.782*****	**Declarative**	0.051	0.387	**1.719*****
Interrogative	0.041	0.624	**3.653*****	**Back-channel**	0.085	0.378	**1.680*****
Back-channel	0.091	0.472	**2.001*****	**Incomplete**	0.003	0.249	ns.

Table 69. Conditional probabilities and odds-ratios of Speaker/Mutual Gaze with Listener Back-channels

Listener Back-channel + Speaker Gaze				**Listener Back-channel + Mutual Gaze**			
	CP | Gaze	**CP | Speech**	**Odds-ratio**		**CP | Gaze**	**CP | Speech**	**Odds-ratio**
Affirmation	0.008	0.622	**3.536****	**Coll. Finish**	0.006	0.507	**2.757*****
Coll. Finish	0.006	0.615	**3.429*****	**Affirmation**	0.006	0.449	**2.188***
Newsmarker	0.004	0.553	**2.652*****	**Acknowledgment**	0.023	0.404	**1.827*****
Acknowledgment	0.023	0.456	**1.811*****	**Newsmarker**	0.003	0.402	**1.795****
Assessment	0.043	0.440	**1.711*****	**Continuer**	0.009	0.385	**1.676****
Continuer	0.008	0.396	**1.405****	**Assessment**	0.038	0.330	**1.329****
Laugh	0.013	0.295	0.893	**Laugh**	0.012	0.231	0.802*

Between speaker gaze and mutual gaze, different listener speech turns show different likelihood patterns. All speaker speech turns are more likely than expected during speaker gaze, with incompletes and declaratives the most common. During mutual gaze, however, all odd-ratios decrease; incompletes are at chance, and interrogatives are the most likely.

For listener back-channels, mutual gaze makes very little change in the likelihoods of overlap but, as with speech turns, all odds-ratios are slightly lower than with speaker gaze. In both overlaps, laughs are the only back-channel type to be less likely than expected.

Table 70 shows the same information for speaker speech turn types and back-channels overlapping with listener gaze and mutual gaze.

Table 70. Conditional probabilities and odds-ratios of Listener/Mutual Gaze with Speaker Speech-types

Speaker Speech turn-types + Listener Gaze				**Speaker Speech + Mutual Gaze**			
	CP | Gaze	**CP | Speech**	**Odds-ratio**		**CP | Gaze**	**CP | Speech**	**Odds-ratio**
Declarative	0.537	0.889	**1.895*****	**Back-channel**	0.014	0.418	**1.934*****
Incomplete	0.094	0.879	**1.318****	**Interrogative**	0.137	0.351	**1.513****
Interrogative	0.108	0.860	**1.103***	**Declarative**	0.580	0.308	**1.449****
Filler	0.026	0.830	0.862**	**Incomplete**	0.060	0.181	0.563*
Back-channel	0.008	0.713	0.436***	**Filler**	0.007	0.066	0.183***

Table 71. Conditional probabilities and odds-ratios of Listener/Mutual Gaze with Speaker Back-channels

Speaker Speech + Listener Gaze				**Speaker Speech + Mutual Gaze**			
	CP | Gaze	**CP | Speech**	**Odds-ratio**		**CP | Gaze**	**CP | Speech**	**Odds-ratio**
Affirmation	0.005	0.668	0.354***	**Affirmation**	0.010	0.420	**1.943*****
Laugh	0.002	0.460	0.150***	**Laugh**	0.002	0.123	0.374***

There are more striking differences between listener gaze and mutual gaze for speaker speech turn types. During listener gaze, back-channels are very unlikely to overlap with gaze, incompletes are more likely, and interrogatives are barely likely (1.10). During mutual gaze, back-channels are the most likely to overlap, incompletes are less likely, and interrogatives are more likely. Fillers are less likely in both cases. The increase in likelihood for back-channels during mutual gaze comes from the fact that gaze tends to overlap when exchanging turns, and speaker back-channels always occur after a listener turn. The increase in likelihood for interrogatives is interesting – interrogatives are appeals for response, and looking at the listener allows them to detect a visual response.

For incompletes, the decrease in likelihood during mutual gaze may suggest that speakers are less interested in detecting a listener response because they have not reached a point in the clause where they feel one is necessary (or likely).

For speaker back-channels laughs remain unlikely across listener gaze and mutual gaze, but, as predicted, speaker affirmations are more likely than expected during mutual gaze.

5.2 N-grams

We now turn to the sequential patterns of gaze and speech. Table 72 shows the bigram pairs of these behavior boundaries.

Table 72. Speaker and Listener Gaze and Speech bigrams (1-second window)

Bigram	Frequency	Symm. CP	CP: 1|2	CP: 2|1
S. Gaze-away + L. Speech offsets	116	0.021	0.173	0.123
L. Speech onsets + **S. Gaze-away**	114	0.020	0.121	0.167
L. Speech offsets + **S. Gaze-away**	101	0.016	0.107	0.150
S. Speech onsets + L. Gaze-towards	99	0.012	0.222	0.054
S. Gaze-towards + L. Speech onsets	88	0.012	0.129	0.091
S. Gaze-away + L. Speech onsets	86	0.011	0.126	0.091
S. Speech offsets + L. Gaze-away	92	0.011	0.219	0.050
L. Gaze-away + **S. Speech offsets**	87	0.010	0.047	0.207
L. Gaze-towards + **S. Speech offsets**	86	0.009	0.047	0.193
S. Speech onsets + L. Gaze-away	71	0.007	0.169	0.039
L. Gaze-towards + **S. Speech onsets**	65	0.005	0.035	0.146
L. Speech offsets + **S. Gaze-towards**	54	0.004	0.056	0.080
L. Gaze-away + **S. Speech onsets**	53	0.004	0.029	0.126
S. Speech offsets + L. Gaze-towards	48	0.003	0.108	0.026
L. Speech onsets + **S. Gaze-towards**	30	0.001	0.031	0.044
S. Gaze-towards + L. Speech offsets	26	0.001	0.039	0.027

We know from Chapter 5 that speech and head often co-occur, and we know from sections 2 and 4 in this chapter that they often co-occur near interlocutor gaze-towards. With that in mind, the findings from Table 72 are not particularly surprising, as they are quite similar to the patterns shown in Table 66 in Section 4, of bigrams of gaze and head boundaries. The major difference is that for gaze and head, the second most frequent bigram is speaker gaze-towards followed by listener head onset. Here, speaker gaze-towards followed by listener speech onset is relatively not quite as frequent. Additionally, there is less of a Zipfian curve to the frequency rankings, with many bigram pairs being similar in frequency.

5.3 Window Histograms

Window histograms of speech and gaze show some patterns, but nothing independent of the closer link between head gesture and gaze.

6. Gaze+Hands

This is the smallest section of chapter 7, as it examines the relationship between the two modalities that have undergone the fewest coding subdivisions.

6.1 Likelihood Measures

Gaze and hands are binarily coded in this dataset, and there are no further likelihood measures beyond those shown in Section 2.1.

6.2 N-grams

Table 73 shows the most frequent bigram pairs of gaze and manual gesture boundaries that occur within a 1-second window of each other.

Table 73. Speaker and Listener Gaze and Hand boundary bigrams (1-second window)

Bigram	Frequency	Symm. CP	CP: 1|2	CP: 2|1
S. Hands offsets + L. Gaze-away	57	0.009	0.132	0.065
S. Hands onsets + L. Gaze-towards	52	0.007	0.119	0.060
L. Gaze-away + **S. Hands offsets**	38	0.004	0.044	0.088
S. Hands onsets + L. Gaze-away	37	0.004	0.086	0.043
L. Gaze-towards + **S. Hands onsets**	37	0.004	0.043	0.084
L. Gaze-towards + **S. Hands offsets**	37	0.004	0.042	0.084
L. Gaze-away + **S. Hands onsets**	35	0.003	0.040	0.081
S. Hands offsets + L. Gaze-towards	27	0.002	0.062	0.031
L. Hands onsets + **S. Gaze-towards**	17	0.004	0.018	0.254
L. Hands offsets + **S. Gaze-away**	13	0.003	0.013	0.206
S. Gaze-towards + L. Hands offsets	12	0.002	0.190	0.012
S. Gaze-away + L. Hands offsets	10	0.002	0.159	0.010
S. Gaze-away + L. Hands onsets	9	0.001	0.134	0.009
S. Gaze-towards + L. Hands onsets	8	0.001	0.119	0.008
L. Hands onsets + **S. Gaze-away**	6	0.001	0.006	0.090
L. Hands offsets + **S. Gaze-towards**	4	0.000	0.004	0.063

Several of the most frequent gaze and manual gesture bigram pairs are quite different from what we saw with gaze and head gesture (the other visible gesture modality). With gaze and head gesture (Section 4.2), it seemed that the most frequent bigrams involved listener head gestures being timed relative to speaker gaze shift. Here, the most frequent bigrams seem to involve listener gaze being timed relative to listener manual gesture boundaries. In the most frequent bigrams, speakers finish their manual gesture and listeners look away, and in the second most frequent, listeners start a manual gesture and listeners look towards.

It is also common, but less frequent, to see speaker manual gesture follow listener gaze-towards, or speaker manual gesture to finish after the listener has looked away. Nevertheless, it seems that listener gaze and speaker manual gesture can each precede or follow each other.

6.3 Window Histograms

There are no clear patterns between hands and gaze in the window histograms.

7. Summary and Hypotheses

7.1 Summary

In this chapter, we looked at the timing relations between behaviors of different modalities, across speakers and listeners. Section 2 looked broadly at all four modalities, without distinguishing by speech or head subtypes. We looked at conditional probabilities and odds-ratios and saw that, of all across-role, across modality pairs, gaze was the strongest predictor of overlap. All listener behaviors were more likely than expected to overlap with speaker gaze-towards, and all speaker behaviors were more likely than expected to overlap with listener gaze-towards. However, when speaker and listener gaze overlaps were taken into account, different modalities behaved differently. During mutual gaze, speaker head gestures became more likely, while speaker manual gesture and speech became less likely (although still more likely than expected). Mutual gaze was the only condition where speaker head gesture was more likely than expected, suggesting that many speaker head gestures are produced communicatively, with the goal of eliciting a response, and seeing the response. We also saw precise timing between listener head gesture onsets and speaker gaze-shifts towards the listener, and much less precise timing

between speaker head gesture onsets and listener gaze-towards, whether because listeners care more or because listener gaze-shift is much less frequent.

In Section 3, we looked in detail at head and speech subtypes. There were several interesting findings with respect to speaker speech types and listener head types. We saw a greater likelihood of towards-motion tilts from listeners during speaker declaratives, and a greater likelihood of listener retractions and juts during speaker interrogatives. We also saw two cases in which listener head gesture patterns mirrored speaker head gesture patterns. During speaker fillers, we saw in Chapter 5 that the most common co-speech head gestures are half-cycle repositioning gestures, and we saw that these are also the most common listener head gestures during speaker fillers. And during speaker affirmations, both nods and shakes of the head are more likely than expected to overlap, as is also the case with listener nods and shakes.

In Section 4, we saw that of all listener head gestures, the nod down was the only one to be less likely than expected during mutual gaze, suggesting it may be a form of nod that is more informative (of comprehension) than communicative. We also saw that the dominant sequence of gaze and head involved speakers looking at listeners, then listeners beginning a head gesture, then speakers looking away, then listeners completing the head gesture. Listeners seem to time their head gesture to follow speaker gaze-away, suggesting that speaker gaze-away may serve to feed information forward to the listener, a ‘front-channel,’ of sorts.

In Section 5, we saw that speaker interrogatives were more likely during mutual gaze than merely during listener gaze, suggesting that not only do speakers use interrogatives to appeal for a response, but they care enough to look for it. In Section 6, we saw that the

most frequent bigram pairs involved listener gaze-shifts being timed relative to speaker manual gesture, as though the gesture attracted the gaze. This is in contrast to bigrams of gaze and head gesture (the other visible gesture modality), in which the most frequent bigrams involved listener head gesture being timed relative to speaker gaze.

7.2 Hypotheses

There are a number of hypotheses one could formulate from the data in this chapter. A small sample is laid out below.

1. Speaker head gestures (overall) are only more likely than expected during mutual gaze, when speaker and listener can see other. This suggests that they are designed to elicit a response. Do speaker head gesture produced during mutual gaze elicit listener responses at a greater rate than speaker head gestures produced during other gaze combinations? (The same could be asked of manual gesture.)
2. Some of the most frequent head and gaze bigram pairs occur when speakers gaze towards and listener nod. The most frequent occur at the end of this: speakers gazing away, followed by listeners completing a head gesture. Why are these pairs so common (and more common than gaze-towards + head onset)? Could it be because speakers time their gestures to end after speakers gaze-away? One could look at correlation between the durations of these head nods and the durations of the gaze-towards.
3. We know about the close timing of listener head gesture and speech onsets to speaker gaze-towards, and we know about the close timing of listener speech onsets to co-speech listener head gestures. Do listener head onsets and speech

onsets show the same timing relationship with speaker gaze-towards when they are produced singly to when they are produced together?

4. If listener head nods can be indicative of attention and/or comprehension, does their timing relative to the moment of window-availability (when the speaker gazes towards them) tell us something about either of these functions? For comprehension, one could manipulate the difficulty of the speaker's message, and for attention one could manipulate the monotony of the message (and compare to control conditions that manipulate general cognitive load), then measure the lag time between speaker gaze-towards and listener head onsets, predicting that one would generate longer lag times than the other for different nod types.
5. Speech turns and back-channels differ on a number of dimensions, including duration, complexity (semantic, syntactic, and pragmatic), and how isolated or uninterrupted they are. This is not precisely a hypothesis, but could one create a hierarchy of 'turniness' based on these dimensions, for different speech types, and possibly also non-verbal articulations, then use these to predict interactions with other behaviors?
6. There is a greater chance of head gesture during speaker declaratives and interrogatives for *both* speakers and listeners (and a lesser chance during speaker fillers and incompletes). This suggests two possibilities: 1) that listeners gesture at a greater rate because of the nature of the speech content in declaratives and interrogatives, or 2) because they are responding to the greater rate of speaker head gestures. Test these possibilities by manipulating the content of the message

and the rate of head gestures (or searching in more qualitative detail at variation within an existing multimodal corpus).

7. We know that certain head gestures are more likely during mutual gaze than merely during interlocutor gaze, and we know that gaze-towards is a strong predictor of interlocutor head gesture. Is mutual gaze-towards an even stronger predictor of interlocutor head gesture?
8. We have suggested that nods down are more likely than other listener head nods to be symptoms of comprehension rather than communicators. Test this by counting rates of different nod types in conversations where participants can see each other and where they can't.
9. We have seen that there is a longer lag time in which listeners begin to nod after speaker gaze towards than the time in which they begin to speak after speaker speech offset. Is this difference in lag time due to the length of speaker glances and speech pauses? Look at speakers with different average durations of glances and speech pauses, and see whether listeners adapt their lag times accordingly.

CHAPTER VIII: CONCLUSIONS AND FUTURE WORK

The two goals of this dissertation have been to 1) identify timing patterns in multimodal interaction, with the goal of helping build more holistic models of human communication, and 2) demonstrate the validity of a particular approach to multimodal analysis: creating a corpus, selecting a set of analytical methods, applying them to every possible combination that has been coded in the corpus, and examining all the results to see which kinds of timing relations stand out from the rest. The purpose of this approach is not to test hypotheses, but to explore new data and formulate hypotheses.

To accomplish these goals, I created a corpus of multimodal communication, coded for head gesture, gaze, manual gesture, and speech. I then ran a set of analyses on every single and pair of behaviors in the corpus, including duration and frequency analyses, measures of likelihood such as conditional probability and odds ratios between observed and expected distributions, and examinations of frequency distributions.

In doing so, I identified a number of patterns related to multimodal timing that I would never have thought to examine

Looking first at within-role, within modality temporal patterns (Chapter 4), I found that the many of the most dependent sequences of speech and head behaviors were repeated bigrams (e.g. for head gestures: single nod + single nod, multiple shake + multiple shake; and for back-channels: assessment + assessment, acknowledgment + acknowledgment), suggesting that speakers and listeners both tend to fall into repeated patterns of production in these two modalities. There were also patterns in the speech and head modalities that reflected the narrative structure of the storytelling itself. Normalizing all

stories to a standard length, there was a robust finding that some behaviors (listener assessments and head shakes) increase in frequency as the rising action of the story increased, and peaked during the climax of the story. This was not the case for other listener head behaviors, such as nods, which remained at a constant frequency throughout the story, providing evidence that listener head gestures do not have a monolithic function. In the opposite direction, listener continuers tended to be used frequently at the beginning of the story, and decreased in frequency as the story approached its climax.

Looking at within-role, across-modality temporal patterns (Chapter 5), I found a number of unexpected trends in the co-occurrence of different modalities with speakers and listeners. Speakers were more likely than chance to produce certain kinds of head gestures while looking at the listener, particularly multiple iterations of nods, shakes, and wags, as well as juts and retractions (behaviors that are directed towards or away from the interlocutor). For listeners, on the other hand, the only head gesture that was more likely to be produced while looking at the speaker was the multiple nod, which is the listener's most frequent gesture. There were also strong temporal tendencies in the co-production of speaker speech types and speaker head gestures. During declarative speech segments, speakers were more likely than expected to produce almost all kinds of head gesture (but especially shakes and wags). During interrogative segments, they were more likely to produce 'motion-towards' gestures and less likely than expected to produce 'motion-away' gestures. During filler speech segments, speakers weren't more likely than expected to produce any head gestures, but the gestures they produced most were half-cycle gestures such as tilts, juts in, or nods down.

Looking at across-modality, within-role temporal patterns (Chapter 6), I found that the degree of overlap between each modality across roles was quite small, with the exception of mutual gaze, which occurred during 27% of the total corpus. However, these periods of mutual gaze were relatively short, often only a second or two in length. A frequent 4-gram of gaze shift involved the speaker looking at the listener, the listener holding the mutual gaze for a moment, then looking away, then the speaker looking away, and finally the listener returning their gaze to the speaker. Long periods of mutual gaze were uncommon. Listener speech onsets followed speaker speech offsets with extremely precise timing (typically within 250ms), and speaker speech onsets followed listener speech offsets with much less precise timing, suggesting that listeners take more care not to interrupt their interlocutor in this kind of communicative context. Head gesture onsets were timed to occur near interlocutor head gesture offsets with less precision, but very reliably, and speakers seemed to follow listeners nearly as much as listeners followed speakers, often making it unclear who was initiating the gestural interchange (something that was always very clear in the speech interchanges).

Looking at across-role, across-modality temporal relations (Chapter 7), I found that listeners timed their head gestures to occur just after the speaker looked at them. The precision of this timing was not quite as precise as listener speech onsets following speaker speech offsets (typically within 500ms), but the window of availability afforded by speaker gaze-towards is longer than the window of availability afforded by a pause in the speaker's speech. Another finding suggested that some listener head gestures might be likely to have the function of signaling information to the speaker, and some more likely to be symptomatic of processing speaker speech. Of all four types of listener head

nod (single, multiple, up, and down), only one was less likely than expected to occur while being looked at: the nod down. All other listener head gestures are more likely than expected to occur while being looked at, suggesting that they are aimed to signal information back to the speaker, while the nod down is done regardless of speaker gaze.

I also identified a number of hypotheses based on the findings that I or other researchers could test on future datasets (see Appendix A, or the final sections of Chapter 4-7). These outcomes suggest that the approach was indeed validated.

This approach is only a beginning, however. While examination of four modalities, within and across role, is indeed a feat, this kind of big data approach to analysis can in principle be applied to any number of modalities, and it is clear that there is a great deal of verbal and non-verbal behavior that is missing from this dissertation, and which might help explain many of the uncertainties in the previous chapters.

For non-verbal behavior, there are several modalities that I would like to add. Blinks have been shown to interact with other non-verbal behaviors, such as head nods (Hömke 2017). Posture is also a key component of non-verbal interaction (Bressem 2013). Facial expressions are also clearly an important factor, and some of these might profitably be added to the coding scheme, such as eyebrow raises and smiles. There are also non-categorical measures of head gesture that I would like to add. Motion-capture of hands, arms, and heads would allow for a more etic approach to identifying the relevant categories of non-verbal behavior.

For verbal behavior, the text of these stories has already been transcribed and phonemically time-aligned using the Montreal Forced-Aligner (McAuliffe et al. 2017). In

addition, it would be extremely helpful to code some measure of pitch contour or prominence, as well as intensity. Having seen how speech and gesture categories interact with the structure of the narration, there is good reason to think that these measures might also correlate with rising action.

There are also a number of interesting gender differences we see in this data set, which did not fit in the dissertation.

My primary goal in the immediate future, however, is not to add to the coding scheme or to continue to subdivide it, but to create one or more corpora to compare these data with. This is important for several reasons. First, given the novelty of using big data analytical methods on multimodal data, it is important to replicate the analyses on an independent dataset. Second, this corpus was created from a fairly homogeneous sample of native American English speaking college students. If one wanted to make general claims about the nature of multimodal communication systems, these same methods would have to be tested on corpora created from other samples of people, who might differ in age, in language variety and culture, or in the kind of communicative context, such as conversation or argumentation. Third, it is problematic to test the hypotheses generated from a particular dataset on the same dataset. This replication is especially important given the nature of these analyses, which are post-hoc to the extreme – we can be sure that a substantial number of the timing patterns found in these chapters are only there by chance, and may not appear in other, similar datasets.

Finally, creating a corpus of the same kind of storytelling dyads in a different language would make it possible to begin to look at what is universal and what is culturally variable in multimodal communication. To date, there are no large, cross-linguistic

studies of multimodal corpora (and, indeed, too few cultures represented in multimodal analysis), and we do not know which patterns are the result of basic human physiology or shared environmental context, and which are conventionalized patterns specific to a given culture. But it is clear to anyone who has interacted with people from another culture that there are differences in when and how to look at someone when they are talking to you, or how large and how frequent your nods should be in response to a speaker.

Some findings of universal characteristics would have important implications for models of production, both for speech and for more holistic models. For example, if the finding in Chapter 6.2.3 that head nods tend to precede accompanying speech holds true across many cultures, this would become something for researchers to pay attention to in designing a model that accounts for the production of multimodal constructions. Gaze might be a universal cue that someone is paying attention to a speaker, although cultures may differ in how acceptable it is to display this attention. And some behaviors may well be culturally variable, such as the specific head gestures that are used to express sympathy or surprise, although I predict that the axis between the speaker and the listener will be used reliably across cultures as a way of communicating interactional pragmatic information.

I do hope this approach and these findings will be of use to other multimodal researchers.

APPENDIX

This appendix aggregates all the hypotheses from the final sections of chapter 4 through 7.

Chapter 4: Within-role, Within-modality

1. There may be many reasons why head gesture exhibits the greatest similarity across roles, of all modalities examined here. 1) Listener frequency approaches speaker frequency because head gesture doesn't impede the speaker's message like listener speech would; 2) the 'window of opportunity' for head gesture (interlocutor gaze) is longer, overall, than the 'window of opportunity' for speech (interlocutor speech pause); or 3) listener head gesture responds to or is somehow dependent on speaker head gesture.
    - The first explanation would be difficult to test directly (we cannot manipulate how much one person's head gesture impedes another's).
    - The hypothesis for the second explanation is that people produce more head gestures when they are easily detectable (i.e. easily seen). This could be tested by comparing individuals who had longer and shorter total gaze-time towards their interlocutors, and looking for a positive correlation between longer gaze-times and greater rate of interlocutor head gesture. The same could be done for speech, in fact, comparing individuals with longer or shorter proportions of speech-time during the story, and correlating these with proportions of the interlocutor's spoken back-channels.
    - The hypothesis for the third explanation is that one interlocutor's head gesture is dependent on the other's. This could be tested by looking at correlations of

proportions of each interlocutor's head gesture, with a positive correlation being evidence for this hypothesis.

2. The average lag time between listener speech segments was 50% longer than the average lag time between listener head segments. Does the rate one of these behaviors contribute more than the other to an onlooker's perception of listener comprehension, rapport, or responsiveness, and would changes in the rate of each modality shift these perceptions equally?
    - One set of hypotheses is that a participant's rate of speech (or head gesture) influences onlookers' judgment of that participant's level of comprehension, rapport, or responsiveness. This could be tested by selecting two sets of video clips from storytelling elicitations: one set with a higher rate of speech or head gesture, and one with a lower rate. Experimental participants would watch the video clips and rate the speakers and listeners in these clips on these three target judgments. A prediction might be that a greater rate of speech and gesture would correlate with higher ratings on all these judgments for both speakers and listeners.
3. A nod followed by a half-cycle repositioning head gesture is a highly frequent bigram for listeners. Do these bigram constructions differ from a solitary listener nod in terms of the content of the speaker speech it co-occurs with?
    - We might hypothesize that the nod + half-cycle bigrams will be more likely than the unigram nods to occur near specific linguistic features, such as clause breaks, discourse connectors, or filled pauses, and test this by finding the odds ratio (see Chapter 2.6) of the overlap of these n-grams with the target speech

segments to see whether these overlaps are more or less likely than expected, and comparing them for the bigrams and unigrams. We could also look at discourse pragmatic features, such as the introduction of new events or referents, or the use of perspective-shifting linguistic forms (e.g. *however, on the other hand*, etc.). Greater than likely overlap between the bigrams and such linguistic segments is one way to provide evidence for a functional interpretation of such a head gesture construction.

4. The most dependent listener back-channel bigrams are repeated back-channels (such as assessment + assessment and acknowledgment + acknowledgment). In these bigrams, is each unit responding to the same topic, or to different topics? It's not entirely clear when and why listeners produce back-channels. Certainly they often respond to the content immediately prior, but are they also influenced by the entirety of the content since their last back-channel, or by the nature of their own last back-channel?
    - The fact that the three most dependent listener back-channel bigrams are repetitions suggests two possibilities: 1) the content of the speaker's speech immediately prior to the back-channels is similar in nature (and so elicits a similar back-channel), or 2) the second back-channel unit in these bigrams is motivated by the same considerations as the first, rather than the immediately preceding speaker's speech.
    - Both of these hypotheses could be tested by examining repeated back-channel bigrams in a corpus and looking at the preceding speaker speech for each bigram unit. If, for most bigrams, the preceding speaker content is similar

before each bigram unit, this would suggest that these repetition bigrams are driven by the immediately prior speaker behavior. If, for most bigrams, the first speaker content is appropriate (by some standard) and the second speaker content is not, this would suggest that both back-channels are driven by the first speaker content.

5. Speakers use more interrogatives towards the beginning of their stories. Are these being used in functionally different ways that change over the course of the story?
    - One hypothesis might be that, at the beginning of the story, listeners use interrogatives to establish rapport, or demonstrate that they care about or are invested in the storytelling. Another might be that, at the beginning of the story, there is more that is unknown, and so listeners ask more questions to fill in these information gaps.
    - Either of these factors could be experimentally manipulated, and the proportion of early-story interrogatives could be compared across conditions. For the first hypothesis, the impetus to exhibit investment or establish rapport could be manipulated by comparing strangers' interactions (as in this dataset) vs. friends (whose rapport would already be established), or by having the participants tell multiple alternating stories (so investment would demonstrated more in the earlier stories, although a decline in the need to demonstrate investment might be conflated with a fatigue effect). If early-story interrogatives were equally frequent for both conditions, this would suggest that their function is information-gathering.

- For the second hypothesis, the impetus to fill in information gaps could be manipulated by having one group of participants listen to stories from a stranger that they had already heard or read before (but they are not meant to let the speaker know this), and the other group listen to stories that were novel to them. If early-story interrogatives were equally frequent in both groups, this would suggest that their function is rapport-building or investment-exhibiting.

6. The narrative structure of these stories is quite specific. Would other kinds of stories show the same patterns (such as increased rates of assessments and listener shakes during rising action and climax)?
    - The underlying hypothesis here is that larger discourse structures influence the use of multimodal behaviors.
    - More specific hypotheses could be made about the stability and instability of different interactions between the structure of a communicative context and the use of multimodal behaviors. For instance, one might hypothesize that assessment back-channels will grow more frequent across all communicative contexts (because the urge to show interest increases as the dialogue drags on) or that this is specific to certain kinds of storytelling (because they follow the rising action and climax of the narration). This could be tested by comparing the frequency distributions of assessments in multiple communicative contexts, such as storytelling, instructions-giving, and persuasive arguments (although it may be that non-storytelling contexts also have rising action and climaxes).

7. How dependent or independent of speaker behaviors are back-channels? Will manipulating the visibility of speaker cues influence the rate of back-channels?
    - We can hypothesize that listeners produce back-channels in response to certain speaker cues, such as gaze direction or pitch contours in speech (in addition to the semantic content of the speaker's speech).
    - To test these, we could obscure these cues and compare the timing of listener back-channels in an obscured-cue group and a non-obscured-cue group. This would be relatively easy for gaze – simply put a screen between participants or have them speak with each other remotely. We might predict that head gesture back-channels would be less frequent, but we might predict that the overall rate of back-channels (including spoken back-channels) would not differ. (We could also manipulate whether or not the speaker could see the listener, such as by blindfolding only the listener, to see whether speaker gaze is influential.)
    - For speech, we could use software to resynthesize the speaker's pitch to a constant frequency, eliminating an important prosodic cue to speech offsets (although not syntactic or semantic cues). Here, we might predict a reduction in spoken back-channels, because the window would be less reliable, but not a reduction in back-channels overall.

Chapter 5: Within-role, Across-modality

1. Listener head onsets tend to precede listener speech onsets. Is this equally true, regardless of whether the speaker is gazing towards them or still speaking?
    - Listener head nods tend to respond to speaker gaze shift, while listener speech tends to respond to speaker pauses. But when listeners produce a head nod and

a spoken back-channel together, the corresponding speaker cues are not necessarily also produced together.

- We might hypothesize that, following Growth Point Theory (McNeill 1992), a single source sends signals to the speech and gesture production systems and times them to align in a similar way, regardless of whether or not both the head nod and spoken back-channel are easily detectable – this would predict little variance in timing regardless of whether the speaker was looking at the listener or speaking. We might also predict that the timing between these co-produced behaviors is influenced by the window of availability, by whether or not one or the other of the two behaviors was easily detectable – this would predict that the timing between the listener head nod and speech would vary depending on whether there was a window of availability opening or closing nearby.

2. Laughs and shakes can both be responses to discomforting things, but they are almost never co-produced. Are they responding to different kinds of discomforting things?
    - One hypothesis might be that they respond to the same kinds of things (but are not co-produced, for some reason, possibly because they convey a different kind of response), while another might be that they respond to different categories of discomforting information – this could be tested using the Fisher's Exact Test to examine overlaps. For example, laughs might respond to things that are both discomforting and surprising, while shakes might respond to things that are discomforting and sad.

3. Speaker head gesture overlaps more than expected with gaze-towards (1.52), while speech overlap with gaze-towards is at chance (0.98), and manual gesture is only slightly more likely (1.22). One possibility that could be explored is that some speaker head gesture is intended to elicit a response, and so the speaker looks to check that the signal has been delivered.
    - Here one might make a strong hypothesis that some speaker head gesture is designed to elicit a visible response, and it is more likely to be produced while the speaker is looking at the listener because the speaker wants to see the visible response. To test this, one could look for an effect in the listener. We can count the proportion of listener head gestures that follow speaker head gesture *during* mutual gaze compared to those that follow speaker head gesture *outside* of mutual gaze. If these are more frequent during mutual gaze, this would support the hypothesis.
4. Speaker multiple nods near back-channels and single nods near fillers both show a strong tendency to precede their accompanying speech onset. This is not the case for nods down near interrogatives, which tend to follow the speech onset.
    - This might be a fruitful area for close qualitative analysis, but here are some possible explanations. Listener multiple nods also tend to precede back-channel onsets, so speaker behavior might be obeying the same conventions in this case. When nods and fillers co-occur, both may be serving to facilitate the upcoming speech.
5. Acknowledgments are more likely than expected to overlap with both head shakes and head nods, which are head gestures that tend to be interpreted as having very

different functions. Do these head gestures correlate with acknowledgments of different kinds of speech content (such as positive or negative affect)?

- This could be tested simply by counting the proportion of times acknowledgments followed speaker content that was positive or negative in affect. Other possible categories that might differentiate these responses could be given vs. new information, surprising vs. unsurprising information, or even declarative vs. interrogative statements.

6. Continuers are the only listener back-channel more likely than expected to overlap with listener gaze-towards. These are also the back-channels that have the least to do with comprehension, and more to do with signaling that the speaker should continue. For the other back-channels, it could be that looking away is also a signal that speech is being processed, which would account for the greater overlap with listener gaze-away.

   - One could hypothesize that listener continuers co-occur with listener gaze because the listener is attending more to the available cues in the speaker than their content. This could be tested by examining the responsiveness to cues in continuers compared with other back-channels (e.g. do they co-occur in response to a greater number of cues, or are they more closely timed with certain cues). It could also be tested by manipulating whether or not the listener can see the speaker, and counting the frequency of continuers – if they are less frequent when the listener can't see the speaker, this would suggest they are produced more in response to visible speaker cues.

7. Speaker half-cycle head gestures are the most likely to co-occur with speaker fillers. These gestures often signify a shift in perspective. It's possible that different kinds of half-cycle gestures correspond to different kinds of perspective shifts, which could be examined qualitatively. If so, this would have interesting implications for planning.
    - If could be that different kinds of half-cycle gestures correspond to different kinds of narrative shifts, such as taking a different viewpoint, reconsidering a previous statement, expanding on a previous statement, or moving on to a new topic. Associations between these could be found by looking at the odds ratios of the overlaps of different half-cycle gestures and different categories of narrative shifts.
8. For speaker non-speech segments, which consists of all frames in which no (speaker) speech is occurring, only multiple nods are more likely than expected. This is probably both because it is a common back-channel behavior (and general acknowledgment of an interlocutor behavior), and because it is a thinking behavior, or a behavior one does when one doesn't quite know what else to do.

Chapter 6: Across-role, Within-modality

1. Some theorize that speech production is designed to distribute information optimally over time in a uniform manner (Bell et al. 2003, Aylett & Turk 2004). An alternative theory might be that speakers distribute information with varying degrees of density, for pragmatic reasons. These theories could be tested by examining the amount of information being co-produced on a variety of modalities, including non-verbal modalities such as those in this corpus as well as intra-linguistic modalities.

- If communicative information is distributed uniformly over time, we might hypothesize that a speaker's multimodal information will co-occur with speech information that is unpredictable or otherwise more difficult to process. For example, iconic manual gesture might co-occur with speech to provide redundant information that is hard to process, or to provide additional semantic information.

2. Are other mutual modality segments (such as co-occurring head or manual gesture) also more likely during mutual gaze?
3. Does speech content during mutual gaze (or mutual head gesture) differ qualitatively from speech content during other gaze combinations?
    - We might hypothesize that speech during mutual gaze is, in some way, more 'important,' and so it is produced while the speaker can assess its effect on the listener. This might mean that the information is more salient to the narrative, more surprising, or more calculated to have an emotional impact. These are rather subjective categories, but different passages of the narrative could be subjectively rated on these scales by participants reading the text. If the speech segments that occur during mutual gaze are rated higher on these categories, this would support this hypothesis.
4. Where do 'enclosed' back-channels (back-channel speech or head gesture that begins after a speaker speech onset and ends before the speech offset) occur within the speech segment? Are they responding to different kinds of information than back-channels that occur at the end of speech segments?

- Back-channels often occur following a speaker speech segment, but also often occur during speaker speech. One hypothesis that would be relatively simple to test is that back-channels that occur during speaker speech are more likely than not to occur during *longer* speaker speech segments (either because there is more information to respond to, or because the listener grows tired of waiting for an opening – if it is the latter these might tend to occur more towards the end of long segments).
- Another hypothesis might be that back-channels respond to speech information that may or may not occur at the end of speech segments. For example, some kinds of information that back-channels may respond to are new information, event completions, or uncertainty. If back-channels are more likely than expected following this kind of information when it is both at the end of a speech segment and internal to the speech segment, this would be good evidence to support this hypothesis.

5. We have seen that continuers are more closely tied to interrogatives than declaratives, and suggested that their precise timing relative to interrogative offsets may be because they are responding more to intonation than speech content. Do back-channels respond more to semantic or acoustic information?
   - According to one hypothesis, back-channels might respond more to acoustic information, while according to the other they would respond more to semantic information.
   - To manipulate the amount of semantic information, we could run two groups: in one a listener must respond to a storyteller whose speech is garbled, in

another language, or in some way unintelligible, and in the other the listener responds to a story they understand. If the listener's back-channels (particularly continuers) exhibit the same timing characteristics across groups, this would support the first hypothesis.

- To manipulate the amount of acoustic information, we could have two groups: in one a listener must respond to a storyteller whose pitch has been flattened, and in the other the listener responds to normal storytelling. If the timing characteristics of the back-channels are similar across groups, this would support the second hypothesis

6. Speaker and listener head gesture onsets are more closely tied to interlocutor head offsets than interlocutor head onsets.
    - One hypothesis might hold that participants time the onsets to the offsets, and are able to predict the offset based on the conventional shapes these gestures usually take. Another hypothesis might hold that they time onsets to onsets, and the pattern we see is because the time it takes to plan and produce their head gesture response is similar to the average duration of the head gesture they are responding to.
    - The first hypothesis could be explored by looking at the timing of head gesture onsets to the offsets of head gestures that are nonstandard (and thus unpredictable) in some way, either in shape or duration or some other feature. The second hypothesis could be explored by looking at the variance in the lag between the onset of the first head gesture and the onset of the second – if true, this should be small.

Chapter 7: Across-role, Across-modality

1. Speaker head gestures (overall) are only more likely than expected during mutual gaze, when speaker and listener can see other. This suggests that they are designed to elicit a response. Do speaker head gesture produced during mutual gaze elicit listener responses at a greater rate than speaker head gestures produced during other gaze combinations? (The same could be asked of manual gesture.)
2. Some of the most frequent head and gaze bigram pairs occur when speakers gaze towards and listener nod. The most frequent occur at the end of this: speakers gazing away, followed by listeners completing a head gesture. Why are these pairs so common (and more common than gaze-towards + head onset)? Could it be because speakers time their gestures to end after speakers gaze-away? One could look at correlation between the durations of these head nods and the durations of the gaze-towards.
3. We know about the close timing of listener head gesture and speech onsets to speaker gaze-towards, and we know about the close timing of listener speech onsets to co-speech listener head gestures. Do listener head onsets and speech onsets show the same timing relationship with speaker gaze-towards when they are produced singly to when they are produced together?
4. If listener head nods can be indicative of attention and/or comprehension, does their timing relative to the moment of window-availability (when the speaker gazes towards them) tell us something about either of these functions? For comprehension, one could manipulate the difficulty of the speaker's message, and for attention one could manipulate the monotony of the message (and compare to control conditions that manipulate general cognitive load), then measure the lag

time between speaker gaze-towards and listener head onsets, predicting that one would generate longer lag times than the other for different nod types.

5. Speech turns and back-channels differ on a number of dimensions, including duration, complexity (semantic, syntactic, and pragmatic), and how isolated or uninterrupted they are. This is not precisely a hypothesis, but could one create a hierarchy of 'turniness' based on these dimensions, for different speech types, and possibly also non-verbal articulations, then use these to predict interactions with other behaviors?
6. There is a greater chance of head gesture during speaker declaratives and interrogatives for *both* speakers and listeners (and a lesser chance during speaker fillers and incompletes). This suggests two possibilities: 1) that listeners gesture at a greater rate because of the nature of the speech content in declaratives and interrogatives, or 2) because they are responding to the greater rate of speaker head gestures. Test these possibilities by manipulating the content of the message and the rate of head gestures (or searching in more qualitative detail at variation within an existing multimodal corpus).
7. We know that certain head gestures are more likely during mutual gaze than merely during interlocutor gaze, and we know that gaze-towards is a strong predictor of interlocutor head gesture. Is mutual gaze-towards an even stronger predictor of interlocutor head gesture?
8. We have suggested that nods down are more likely than other listener head nods to be symptoms of comprehension rather than communicators. Test this by

counting rates of different nod types in conversations where participants can see each other and where they can't.

9. We have seen that there is a longer lag time in which listeners begin to nod after speaker gaze towards than the time in which they begin to speak after speaker speech offset. Is this difference in lag time due to the length of speaker glances and speech pauses? Look at speakers with different average durations of glances and speech pauses, and see whether listeners adapt their lag times accordingly.

REFERENCES CITED

Altorfer, A., Jossen, S., Würmle, O., Käsermann, M. L., Foppa, K., & Zimmermann, H. (2000). Measurement and meaning of head movements in everyday face-to-face communicative interaction. *Behavior Research Methods, Instruments, & Computers*, *32*(1), 17-32.

Andersen, P. A. (1999). *Nonverbal communication: Forms and functions*. Mountain View, CA: Mayfield.

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*(1), 28-55.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, *47*(1), 31-56.

Baron-Cohen, S., Baldwin, D. A., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language?. *Child development*, *68*(1), 48-57.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566-580.

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

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001-1024.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive foundations of interpretation* (pp. 69-94). KNAW.

Bressem, J. (2013). A linguistic perspective on the notation of form features in gestures. *Body–language–communication: An international handbook on multimodality in human interaction*, *1*, 1079-1098.

De Ruiter, J. P. (1998). *Gesture and speech production* (Doctoral dissertation, Radboud University Nijmegen Nijmegen).

De Ruiter, J. P. (2000). The production of gesture and speech. *Language and gesture*, *2*, 284.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.

Drummond, K., & Hopper, R. (1993). Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on language and Social Interaction*, *26*(2), 157-177.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, *23*(2), 283.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49-98.

Fromkin, V. A. (1973). Appendix: A sample of speech errors. *Speech errors as linguistic evidence*, 243-269.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance* (Vol. 92). John Benjamins Publishing.

Garrett, M. F. (1988). Processes in language production. *Linguistics: the Cambridge survey*, *3*, 69-96.

Goldin-Meadow, Susan. "The role of gesture in communication and thinking." *Trends in cognitive sciences* 3, no. 11 (1999): 419-429.

Goldman-Eisler, F. (1968). Psycholinguistics: Experiments in spontaneous speech.

Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry*, *50*(3-4), 272-302.

Goodwin, C., & Goodwin, M. H. (1987). Concurrent operations on talk. *IPrA Papers in Pragmatics*, *1*(1), 1-54.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, *2*, 73-113.

Hadar, U., Steiner, T. J., & Rose, F. C. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), 214-228.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts of head postures during conservation. *Human Movement Science*, *3*(3), 237-245.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596-615.

Haviland, J. M. (1977). Sex-Related Pragmatics in Infants Nonverbal Communication. *Journal of Communication*, *27*(2), 80-84.

Heinz, B. (2003). Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, *35*(7), 1113-1142.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Hömke, P., Holler, J., & Levinson, S. C. (2017). Eye blinking as addressee feedback in face-to-face conversation. *Research on Language and Social Interaction*, *50*(1), 54-70.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin*, *137*(2), 297.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Seventh European Conference on Speech Communication and Technology*.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2014). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, *57*, 233-243.

Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental psychology*, *33*(3), 453.

Jefferson, G., Sacks, H., & Schegloff, E. A. (1987). Notes on laughter in the pursuit of intimacy.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and cognitive processes*, *24*(2), 313-334.

Kendall, T., Bresnan, J., & Van Herk, G. (2011). The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*, 22-63.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, *25*(1980), 207-227.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language*, *48*(1), 16-32.

Kousidis, S., Malisz, Z., Wagner, P., & Schlangen, D. (2013). Exploring annotation of head gesture forms in spontaneous human interaction. In *Proceedings of the Tilburg Gesture Meeting (TiGeR 2013)*.

Krauss, C., & Chen, J. Gottesman (2000) Lexical Gestures and Lexical Access: A Process Model. *Language and gesture*, 261-283.

Krauss, R. M., Garlock, C. M., Bricker, P. D., & McMahon, L. E. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of personality and social psychology*, *35*(7), 523.

Labov, W. (1972). *Sociolinguistic patterns* (No. 4). University of Pennsylvania Press.

Leonard, T., & Cummins, F. (2009). Temporal alignment of gesture and speech. *Proceedings of Gespin*, 1-6.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and brain sciences*, *8*(4), 529-539.

Loehr, D. P. (2004). *Gesture and intonation* (Doctoral dissertation, Georgetown University).

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, *3*(1), 71-89.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science*, *36*(8), 1404-1426.

Lucero, C., Zaharchuk, H., & Casasanto, D. (2014, January). Beat gestures facilitate speech production. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of experimental child psychology*, *41*(1), 93-113.

Maynard, S. K. (1987). Interactional functions of a nonverbal sign Head movement in japanese dyadic casual conversation. *Journal of pragmatics*, *11*(5), 589-606.

Maynard, S. K. (1997). Analyzing interactional management in native/non-native English conversation: A case of listener response. *IRAL, International Review of Applied Linguistics in Language Teaching*, *35*(1), 37.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of interspeech* (pp. 498-502).

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of psycholinguistic research*, *23*(1), 45-66.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, *32*(7), 855-878.

McNeill, D. & Duncan, S.D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and Gesture*, pp. 141-161. Cambridge: Cambridge University Press.

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

McNeill, D. (Ed.). (2000). *Language and gesture* (Vol. 2). Cambridge University Press.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615.

Morsella, E., & Krauss, R. M. (2004). The role of gestures in spatial working memory and speech. *The American journal of psychology*, 411-424.

Neidle, C. J., Kegl, J., MacLaughlin, D., Bahan, B., & Lee, R. G. (2000). *Syntax of American Sign Language: Functional Categories and Hierarchical Structure (Language, Speech, and Communication)*. MIT Press.

Novick, D. G., Hansen, B., & Ward, K. (1996, October). Coordinating turn-taking with gaze. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1888-1891). IEEE.

Pine, K. J., Bird, H., & Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. *Developmental Science*, *10*(6), 747-754.

Poggi, I., D'Errico, F., & Vincze, L. (2010, May). Types of Nods. The Polysemy of a Social Signal. In *LREC*.

Query, W. B. I. S. (2016). Reporting System (WISQARS). Centers for Disease Control and Prevention. *National Center for Injury Prevention and Control*.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, *7*(4), 226-231.

Redford, M. A. (2013). A comparative analysis of pausing in child and adult storytelling. *Applied Psycholinguistics*, *34*(3), 569-589.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological science*, *18*(5), 407-413.

Rochet-Capellan, A., Laboissière, R., Galván, A., & Schwartz, J. L. (2008). The speech focus position effect on jaw–finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1507-1521.

Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. In *Speech Prosody 2010-Fifth International Conference*.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). (1974). A simplest systematics for the organization of turn-taking in conversation. Language, 50, 696-735.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, *71*, 93.

Schegloff, E. A. (1987). Analyzing single episodes of interaction: An exercise in conversation analysis. *Social psychology quarterly*, 101-114.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica*, *8*(4), 289-327.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361-382.

Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., & Renwick, M. (2007). A Method for Studying the Time Alignment of Gestures and Prosody in American English:Hits' and Pitch Accents in Academic-Lecture-Style Speech. *NATO SECURITY THROUGH SCIENCE SERIES E HUMAN AND SOCIETAL DYNAMICS*, *18*, 34.

Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*(2), 85-89.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, pnas-0903616106.

Strean, W. B. (2009). Laughter prescription. *Canadian Family Physician*, *55*(10), 965-967.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 1-21.

Thompson, R., Emmorey, K., & Kluender, R. (2006). The relationship between eye gaze and verb agreement in American Sign Language: An eye-tracking study. *Natural Language & Linguistic Theory*, *24*(2), 571-604.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559).

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970* (pp. 567-578).

Zipf, G. K. (1949). Human behaviour and the principle of least-effort. Cambridge MA edn. *Reading: Addison-Wesley*.