GESTURE AND INTONATION

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Daniel P. Loehr, M.S.

Washington, DC
March 19, 2004

GESTURE AND INTONATION

Daniel P. Loehr, M.S.

Thesis Advisor: Elizabeth C. Zsiga, Ph.D.

ABSTRACT

This dissertation investigates the relationship between gesture and intonation. Gesture is known to correlate on a number of levels with speech in general, but less is known about gesture's relationship to intonation specifically.

I filmed four subjects in natural conversations with friends, and annotated sections of the resulting digital videos for intonation and gesture. I annotated intonation following ToBI (Tones and Break Indices), an implementation of Pierrehumbert's (1980; Beckman and Pierrehumbert 1986) intonational framework. I coded gesture according to guidelines published by McNeill (1992) and colleagues. Over 7,500 time-stamped annotations were manually recorded in a digital annotation tool, and exported for statistical analysis.

I sought answers to five questions. First, does Bolinger's (1983, 1986) hypothesis hold, in which pitch and body parts rise and fall together, to reflect increased or decreased tension? I found no evidence of this.

Second, each modality has hypothesized units. Do the unit *boundaries* align? I found that the apexes of gestural strokes and pitch accents aligned consistently, and gestural phrases and intermediate phrases aligned quite often.

Third, do the various unit *types* correlate? I found no significant correlation between movement types (e.g. deictics, beats) and tone types (e.g. pitch accents, phrase tones).

Fourth, do the respective *meanings* of gestural and intonational events correlate? Although intonation is semantically and pragmatically impoverished relative to gesture, I did find occasional but striking instances where the meanings of the two modalities converged.

Finally, how do the two modalities integrate rhythmically? I found a rich relationship, in which the three main "instruments" (hands, head, voice) interplayed much like a jazz piece, with tempos that sometimes synchronized, sometimes differed, and which included full notes, half notes, and syncopation.

The findings are relevant to theories proposed for each modality. For intonation, gestural counterparts to intermediate phrases provide independent evidence for the existence of such phrases. For gesture, the observed relationship with intonation lends support to the theory of a common cognitive origin of gesture and speech.

ACKNOWLEDGEMENTS

It takes a village to raise a child, and it has taken a community of scholars, colleagues, friends and family to nurture this dissertation to fruition.

My thesis advisor, Lisa Zsiga, introduced me to phonology and to her enthusiasm for it. A careful theoretician, she helped ensure that my interdisciplinary study was firmly grounded in intonational theory. Cathy Ball has been my heartily supportive academic advisor for many years. A fellow computational linguist, she encouraged me to look further afield, especially to pragmatics, to better understand how humans might interact with computers. Cathy also taught me the value of conscientious research, through her own example. Ron Scollon helped me understand the nuances of sociolinguistics in general, and interactional studies in particular. His sharp eye noticed at a glance many phenomena in my videos which I had missed in hours of scrutiny.

The broader research community has been especially helpful along the way. David McNeill, Adam Kendon, and Janet Pierrehumbert all gave me early encouragement of my ideas. Keli Yerian provided the inspiration for my hypotheses with her survey paper of gesture and intonation studies. Evelyn McClave, who has also written a dissertation on this topic, was very supportive of my approach. I hope my findings have built on those she first discovered.

There are two scholars in particular without whom I wouldn't have been able to conduct my research. Sue Duncan shared with me her wealth of experience as a "gesturologist", and taught me how to annotate gestures (and the difficulties therein). Through many conversations with Sue, I learned the real issues in the field. I owe a similar debt to Katy Carlson, for sharing her wisdom in intonation, and for patiently helping me learn ToBI annotation. Any mistakes I've made in annotation or theory are not due to these two fine teachers.

I'm extremely grateful to my colleagues at The MITRE Corporation, especially Susann LuperFoy, Linda Van Guilder, Flo Reeder, Christy Doran, Keith Miller, George Wilson, Inderjeet Mani, and too many others to list in the amorphous "language group". I owe the greatest thanks, however, to Lisa Harper, a longtime collaborator and friend, and the one who first got me interested in gesture. My research has been profoundly shaped by our work together. I also thank MITRE for financial support, including the chance to write my dissertation under the Accelerated Graduate Degree Program. Without this assistance, I simply would not have finished.

Michael Kipp made not only this dissertation but a whole field of investigation possible with his development of Anvil. My method of analysis wouldn't have been feasible without Michael's software, and he was cheerfully responsive to my many questions.

I thank the video subjects who trusted me and who happily carried out their assignment to "converse freely and naturally" with their friends. Most of them continued chatting even after the cameras were turned off.

I was supported by many, *many* friends. They really are too numerous to mention, so to avoid omitting any, I will let them remain anonymous. Fortunately, they know who they are.

I thank my father, Raymond Loehr, who earned his Ph.D. when I was born, for always encouraging me to do my best. My mother, Joan Briggs Loehr, has been my advocate beyond even the high standards of motherhood. My seven siblings and their partners have modeled living full yet balanced lives. Susan Rose, Anne Loehr, and Neel Inamdar allowed me to use their homes as a graduate study carrel for many months. My children, Maya and Jake, taught me through their examples both the importance of studying language, and the ultimate unimportance of it compared to more important things in life. (Regarding their examples of language, I once overheard my son, then two years old, arguing with a jay. The bird was making a falling two-

toned caw, and after every caw, my son would reply with a defiant *uh-HUH*! He had obviously interpreted the caw as *UH-uh*, and he was determined to win the cross-species intonational argument.)

My ultimate and greatest appreciation goes to Karen Rose Loehr, my wife, for her unconditional love. Karen has had an unwavering conviction that I could and would create this dissertation. It is to her that I dedicate this work.

TABLE OF CONTENTS

1 Introduction

Language is more than words. Much of the information conveyed in face-to-face conversation is carried in channels other than the lexical stream alone. Two such channels are intonation and gesture. Speakers are aware that intonation carries meaning, sensing, for example, that "questions go up, and statements go down." As for gesture, movements such as deixis (pointing) are inextricable from such lexical counterparts as "that one, over there". But intonation and gesture convey much more than sentence type or deictic reference. Researchers have shown that they each convey a rich set of semantic and pragmatic information.

Relatively recently, a handful of researchers have been exploring the interaction *between* gesture and intonation. The types of interaction studied include temporal alignments of the two modalities, the mapping between their respective hierarchical structures, and a possible shared expression of emotion. More generally, there is strong evidence, provided initially by Kendon (1972, 1980) and McNeill (1985, 1992), that gesture and speech in general are two surface facets of a single underlying thought being expressed.

Of these few interdisciplinary studies, however, only two have been done in depth (Bolinger 1983, 1986; McClave 1991, 1994), and neither of these has used theories of autosegmental intonational phonology (Ladd 1996), of which Pierrehumbert's (1980; Beckman and Pierrehumbert 1986) is an example. Nor has either used acoustic measurements in transcribing intonation. I have chosen, therefore, to examine the relationship between intonation and gesture from an autosegmental intonational phonology framework, while using acoustic measurements.

Such a study has potential importance on at least five fronts. First, it could corroborate or refute existing claims, such as Bolinger's (1983, 1986) claim that pitch and gesture rise and fall in parallel. Second, it could have theoretical implications, providing independent evidence or

counter-evidence to inform the internal debates in each field. In intonation research, for example, Pierrehumbert's intermediate phrases have been debated, while in the gesture community the very function of gesture (whether it assists the speaker or the listener) is in dispute. Third, an investigation along these lines may reveal as yet-unknown correlations between gesture and speech, akin to Condon's (1976) observations that listeners move in rhythm with speakers. Fourth, from a computational perspective, more insight could help researchers build better systems for understanding and generating intonation and gesture in human-computer interfaces. Finally, and most fundamentally, because gesture and intonation are parallel surface forms of deeper thoughts, such research could give us more precise tools to measure the underlying workings of language, within the human brain.

I credit Yerian (1995) for providing the initial idea for this research. As mentioned, the studies linking gesture and intonation are few. Yet several correlations have been found: for instance, an alignment between certain gestural and intonational units. Yet the intonational units used in these studies, things like "nuclei", "tone units", "rises", and "falls", have been challenged in the intonational literature by theories of intonational phonology, of which Pierrehumbert's is one example. Yerian (1995), therefore, proposed re-examining the correlation using Pierrehumbertian intonational units.

Yerian posed a variety of questions (1995, pp. 15-18). For instance, she delves into Bolinger's "Parallel Hypothesis", in which people purportedly raise and lower pitch and body parts together, to reflect increased or decreased tension:

> The proposal that gestural movement may at times parallel the direction of pitch movement may be looked at more closely beyond terminal rises and falls. When there is parallel movement, does it coincide with certain … pitch accents or phrases?

Yerian asked other questions about the relationship between the two modalities:

> Are there patterns of body movement which correspond to … pitch accents, intermediate phrases, and intonational phrases? Are the holds of gestures sensitive to any or all of these boundaries? When more than one gesture occurs in one … intonational phrase … do these gestures show sensitivity to intermediate phrase accents …? … Are [holds] dropped before, during, or after boundary tones? … Might Pierrehumbert's hypothesis that intermediate phrases are characterized by 'catethesis' … find support in some parallel form of gestural movement? … Almost all of the gesture research … reports strong tendencies for accent and gesture to be aligned, but exceptions always exist. Perhaps if different *kinds* of accents … were considered, some additional patterns may emerge. For example, gestures … may align themselves differently in response to the various complex pitch accents Pierrehumbert proposes.

Yerian pointed out that beyond providing specific insights, such questions have theoretical implications:

> Answers to these questions may have a significant impact on … linguistic theory in general.… if we find that gestural units coincide consistently with the phrase units proposed by Pierrehumbert, we have further evidence that these units do indeed exist. Moreover, if the use of gesture in some way mirrors the phenomenon of catethesis, Pierrehumbert would have even stronger evidence that intermediate phrases exist in English.

In other words, clearer evidence from the interaction of the two fields may provide evidence or counterevidence for the units proposed by each. Yerian concluded by pondering even more fundamental questions:

> Perhaps a more important consequence of such findings … would be a deeper understanding of whether intonation and gesture do in fact seem to stem from a single underlying structure … it may illuminate some of the trickier questions of how meaning might be expressed via these mediums, and whether facets of the same meaning are being 'unpacked' differently or similarly in each.… there is little doubt that such an effort would … promote the search for a more integrated and holistic model of communication by all.

Yerian certainly set forth a worthwhile research agenda. However, she did not pursue it herself, focusing her subsequent work on the use of what she terms the “discursive body” (integrating vocal and non-vocal interactional strategies) in women’s self-defense courses (Yerian

2000). I have chosen, therefore, to try and answer the questions Yerian posed, as well as some of my own.

My general hypothesis is the following. A careful analysis of the unit boundaries and types of both gesture and intonation will reveal structural, temporal, and possibly pragmatic parallels, of the type noticed by Kendon and McNeill with regards to other aspects of speech.

More specifically, I ask five questions, the first three inspired by Yerian.

To start with, does Bolinger's "Parallel Hypothesis" hold, in which people raise and lower pitch and body parts together to reflect increased or decreased tension? This addresses the first of Yerian's queries. The rest can be covered by the following two general questions. Given that each modality has hypothesized units, how do their unit *boundaries* and other landmarks align? And how do their various unit *types* correlate?

In addition to Yerian's proposals, I ask two more questions. How do the respective *meanings* of gestural and intonational events correlate? Finally, how do the two modalities integrate rhythmically?

To answer these questions, I filmed four subjects in natural conversations with friends, and annotated sections of the resulting digital videos for both intonation and gesture. Using the Praat audio annotation tool (Boersma 2001), I annotated intonation according to the ToBI (Tones and Break Indices) scheme, an implementation of Janet Pierrehumbert's (1980; Beckman and Pierrehumbert 1986) intonational framework. Using the Anvil video annotation tool (Kipp 2001), I coded gesture according to the guidelines published by David McNeill and colleagues. Over 7,500 time-stamped annotations were manually recorded from 164 seconds of video, comprising nearly 5,000 video frames. The annotations were then exported for statistical analysis. The Anvil tool also permitted a time-aligned view of both intonation and gesture annotations, along with the

raw video and audio. Therefore, using both statistical and observational methods, I was able to answer each question.

This dissertation consists of seven chapters, including this introduction. In Chapter 2, I review the literature of each field, focusing on those works that deal with their interaction. I follow this with a description of an initial pilot study, in Chapter 3. In Chapter 4, I outline my methodology. In Chapter 5, I present my results, and then discuss them in Chapter 6. Chapter 7 contains concluding remarks, and suggestions for future research.

2 Review of Related Literature

As Yerian (1995) pointed out in her excellent overview of intonation/gesture research up to that point, research into the interaction between gesture and speech was first undertaken by gesture researchers, not by linguists. Thus, I will initially (Section 2.1) look at studies from the gesture field. Section 2.2 will then chronicle the work by researchers looking at intonation, as it relates to gesture.

For a number of researchers who touch on both modalities, my decision may seem rather arbitrary as to which section of the literature review I discuss them in. As researchers specifically investigating intonation and gesture are rarer, I have chosen to discuss most of these in my intonation overview. However, pioneers of the gesture field (e.g. Birdwhistell, Kendon, and McNeill) are discussed in my gesture overview, even though they also consider intonation to some extent.

2.1 Gesture

My review of the gesture literature will consist of three parts. First, I will define the type of gesture I am studying. Second, I will discuss earlier gesture research, up to and including the work of David McNeill, who made the importance of gesture widely known with the publication of his 1992 book *Hand and Mind: What Gestures Reveal about Thought*. Following McNeill's book, the field of gesture studies has blossomed. To coherently present the growing body of work since McNeill (1992), I have therefore organized the third section of my gesture literature review along the lines of a debate about the functions and cognitive origins of gesture.

As my dissertation will focus on gestures of the arms, hands, and fingers, and to some extend the head, I will give less attention to studies which focus on facial or torso movements, or the full range of head movements. These movements are equally as important as arm/hand/finger

gestures, yet the enormity of transcribing every movement of the body will require me leave to future research the relationships of such movements to intonation.

For a thorough, book-length overview of nearly all aspects of the gesture-speech relationship (including semiotics, psychology, and child development), see Feyereisen and de Lannoy (1991).

2.1.1 Definition of Gesture

Gesture, as usually defined by the field that studies it, refers to spontaneous bodily movements that accompany speech[1]. The most common body parts used are the hands, fingers, arms, head, face, eyes, eyebrows, and trunk. Gesture does not include movement that does not accompany speech. Nor does it include pantomime, or emblematic gestures such as the "OK" sign in North America, or signed languages such as American Sign Language. Definitions for all these types of movements were provided by Kendon (1982), who prefers the term *gesticulation* for what I (in keeping with most of the literature) will refer to as *gesture*. McNeill (1992, 2000a) arranged Kendon's movement types into a continuum:

Gesticulation → Emblems → Pantomime → Sign Language

There are several dimensions of this continuum (McNeill 2000a). The first has to do with the movement's *relationship to speech*:

Gesticulation	→	Emblems	→	Pantomime	→	Sign Language
Obligatory presence of speech		Optional presence of speech		Obligatory absence of speech		Obligatory absence of speech

[1] The word *gesture* has another meaning in the field of articulatory phonetics, where it is used to describe movements of the speech articulators (e.g. the tongue). I am not studying that kind of gesture.

Gesture (Kendon's *gesticulation*), then, occurs while speaking (and, as will be discussed, is integral to the spoken utterance). On the other end of the spectrum, sign language occurs in the absence of speech, because sign languages have a fully structured linguistic system. This leads us to McNeill's next dimension of the continuum, *relationship to linguistic properties (relationship to conventions)*:

Gesticulation	→	Pantomime	→	Emblems	→	Sign Language
Linguistic properties absent, not conventionalized		Linguistic properties absent, not conventionalized		Some linguistic properties present, partly conventionalized		Linguistic properties present, fully conventionalized

By the "absence of linguistic properties", McNeill means that there is no lexicon of agreed-upon symbols, and no phonological, morphological, and syntactic system of combining any such symbols. (Note that pantomime and emblems have switched places in this dimension). Gesture, in other words, has no conventionalized constraints on how a concept is to be communicated.

Finally, McNeill's last dimension concerns the *character of the semiosis*:

Gesticulation	→	Pantomime	→	Emblems	→	Sign Language
Global and synthetic		Global and analytic		Segmented and synthetic		Segmented and analytic

The sense of *segmented* is that the meaning of the whole is determined by the meaning of the parts, in a bottom-up fashion. In language (spoken or sign), meanings of segmented morphemes are assembled into larger meanings. Conversely, the sense of *global* is that the meaning of the parts is determined by the whole. McNeill's example is based on a subject's gesture accompanying the speech, "he grabs a big oak and bends it way back". During the words "bends it", the subject's hand makes a grasping motion along with a pulling motion back and

down. (It should be noted that the gesture is indistinct enough to not be recognizable as pantomime.) The 'parts' of this particular gesture could be thought of as the hand, its movement, and the direction of movement. "These are not independent morphemes. It is not the case that the hand in general means a hand or movement backward must always mean movement in that direction …" (McNeill 2000a, p. 5). Rather, the fact that the movement backwards meant *pulling* in this case is determined by the *whole* gesture's enactment of bending back a tree. Thus, the semiosis of gesture is global.

The sense of *synthetic* is that a gesture combines into one symbol disparate meanings, which must be divided across the words of the accompanying sentence. The meanings of actor, action, and manner, which the sentence captured in separate words (*he, bends, back*), were displayed in a single gesture. The sentence is conversely *analytic*; its meaning is divided across symbols (successive words).

Put another way, *global-vs.-segmented* deals with how meaning is interpreted (top-down or bottom-up). *Synthetic-vs.-analytic* deals with how meaning is distributed across symbols (one symbol or multiple symbols).

In a *global* gesture, the meaning of the parts is determined from the meaning of the whole, top-down. In a *segmented* sentence, the meaning of the whole is determined from the meaning of the parts, bottom-up.

In a *synthetic* gesture, the meaning is conveyed within a single symbol. In an *analytic* sentence, the meaning is conveyed across multiple symbols (words).

Some examples may help clarify McNeill's analysis of Kendon's continuum. Let's start with "pure" gesture, Kendon's "gesticulation". Imagine watching a television talk-show host with the sound turned off. While speaking, the host makes movements with her hands. The chances are, without the benefit of sound, that the observer will have no idea what the movements

mean. This illustrates gesture's obligatory presence of speech—the movement can not be interpreted without the accompanying words. The reason it can't be interpreted is because gesture has no conventionalized or agreed-upon inventory of linguistic symbols. Instead, gesture is composed of arbitrary, spontaneous, and unconstrained movements.

Let's say we now turn the sound on for a while, and watch and hear the host talk about home construction. She makes a movement with her hands grasping an imaginary object and moving it down. Thanks to the accompanying speech track, we know the meaning is "stacking bricks". This movement has in fact two meanings: one of *bricks*, and one of *stacking*. The accompanying speech has a symbol for each meaning: the symbol (word) *bricks*, and the symbol (word) *stacking*. The semiosis of speech is therefore analytic. Yet in the gesture, the multiple meanings are distributed within a single symbol. The semiosis of gesture is therefore synthetic.

Finally, the gesture will be composed of different elements—for example, one element of hand shape (fingers mostly closed), and another element of hand direction (moving down). Such a hand shape (fingers mostly closed) has no meaning on its own. Neither does such a hand direction (moving down). Rather, the meaning of each of these elements is interpreted from the meaning of the whole gesture. Thus we know that—in this situation—*fingers-mostly-closed* means grasping a brick, and *hands-moving-down* means stacking. Because the elements of a gesture get their meaning from the meaning of the gesture as a whole, then the semiosis of a gesture is global (top-down meaning).

Let's look at the opposite end of the spectrum: sign language. Sign language needs no speech to be understood because it is understandable in isolation. This, in turn, is because the movements are conventionalized, agreed-upon linguistic symbols. Meaning is distributed across a series of symbols, not a single one; hence sign language is analytic. And in sign language,

elements such as hand form and direction *are* morphemes in their own right which contribute to the meaning of the sign; hence a sign is segmented (bottom-up meaning).

How about pantomime? As in gesture, the movements in pantomime are not a set of agreed-upon, conventionalized symbols. Yet their meaning is recognizable to the audience thanks to the skill of the mime's acting. Therefore, like sign language, pantomime is meant to occur in the absence of speech. Also like sign language, the overall meaning is distributed across individual movements—each movement is almost a stand-in for a word. Hence, pantomime is analytic. But unlike in sign language, elements such as hand form or hand direction have no meaning on their own. Their meaning is interpreted in light of the movement as a whole. Thus, a movement in pantomime has global semiosis (top-down meaning).

The last element in the continuum is the emblem. As an example, let's use the North American "OK" emblem, with thumb and index finger in a circle, and the other fingers extended. This can be understood in the absence of speech, though it's allowed to be used in conjunction with speech as well. It's conventionalized and has some, but not all, linguistic properties. For example, it modifies something as being "OK", but the noun or situation being modified can be outside of the emblem system. The meaning of "OK" emblem is distributed across a single symbol; hence, it's synthetic. Finally, McNeill labels emblems as having segmented semiosis (bottom-up meaning), presumably because multiple emblems can be strung together to create larger meanings.

The main point to take away from McNeill's descriptions is the following. Language (speech or sign) is serial, conventionalized, and self-supporting with regards to meaning. Gesture is holistic (including space as well as time), arbitrary, and dependent upon speech for meaning.

2.1.2 Earlier Studies of Gesture

Having defined what gesture is and is not, I'll now review the literature of gesture studies. As mentioned, I will first review earlier works, up to and including McNeill (1992). This will include discussion of the various forms of gesture that researchers have identified, including the categories they have classified gestures into. Though terminology differs (see Rimé and Schiaratura (1991) for an excellent compilation and cross-reference), there is little dissension as to the basic types.

2.1.2.1 Classical Studies

From a rhetorical perspective, gestures have been studied as far back as the ancient Indians. Hindu treatises on the pronunciation of the Sanskrit Vedas prescribed not only tonal accents (high, low, and rising-falling), but also manual gestures to accompany these accents (near head, heart, and ear, respectively). Thus, the height of the gesture matched the "height" of the tone (Shukla 1996, pp. 274-275). In the first century A.D., the Roman Quintillianus prescribed the gestures orators should use during speech-making (Quintillianus 1979). Quintillianus coined the term *chironomia*, or "rule of the hand", to refer to this manual art of rhetoric.

During the Renaissance, Francis Bacon observed a relationship between the mind and gesture, and not just between the mind and speech. "Aristotle sayth well", wrote Bacon, "Wordes are the Images of Cogitations … But yet [it] is not of necessitie, that Cogitations bee expressed by the Medium of Wordes. For … mens minds are expressed in gestures." (Bacon 1605/2000, pp. 119-120). Bulwer (1644/1974) was inspired by both Quintillianus' prescriptivism and Bacon's descriptivism. He wrote a treatise discussing not only Quintillianus' chironomia, but also on what Bulwer termed *chirologia*, or "language of the hand", examining the meanings of gesture. During the Enlightenment, Condillac (1792/1982, 1746/1973, pp. 194-199) was representative of a number of philosophers who speculated that human language originated with gestures.

In more modern times, Charles Darwin (1872/1998) strove to show that bodily expressions of emotion in humans and certain animal species were similar, innate, and universal, thereby supporting his theory of evolution (Ekman 1998). Edward Sapir noted that "… we respond to gestures with an extreme alertness and, one might almost say, in accordance with an elaborate and secret code that is written nowhere, known by none, and understood by all" (Sapir 1927/1974, p. 42). In connection with his theories on language and culture, Sapir felt that this unwritten code of gesture was a product of social tradition. The psychologist Wundt (1921/1973) theorized about the cognitive origins of gestures, crediting their source to expressive emotions. Like others before him, Wundt felt that gesture is an "expression of thought" (p. 55), and that "gestural communication is a natural product of the development of expressive emotions" (p. 127). The psychologist Wolff (1945/1972) had similar theories, thinking of gesture as a personality indicator akin to a Rohrschach (ink-blot) test. Bloomfield (1933), in a point relevant to the present work, linked gesture and intonation together, due to the paralinguistic qualities of each. He wrote (p. 114), "we use features of pitch very largely in the manner of gestures, as when we talk harshly, sneeringly, petulantly, caressingly, cheerfully, and so on."

2.1.2.2 Efron

The first to carefully study spontaneous speech-accompanying gestures, however, was Efron (1941/1972), who studied the hand movements of New York City immigrant populations. He refuted Nazi claims that gesture was inherited by "race", by showing that while non-assimilated Jews and Italians gestured differently from each other, their assimilated children gestured similarly to each other, and not like their forbears. Thus, culture was the deciding factor.

Efron broke methodological ground for describing gestures, and came up with the following classification[2]. (All quotes, unless otherwise specified, are from Efron 1941/1972, p. 96).

Efron proposed two main categories of gesture, depending on whether the gesture had meaning independent of speech, or in conjunction with speech.

The gestures with meaning independent of speech he labeled *objective* gestures. These are meaningful by the "connotation (whether deictic, pictorial, or symbolic) it possesses independently from the speech of which it may, or may not, be an adjunct". There are three subcategories. A *deictic* refers "by means of a sign to a visually present object (actual pointing)". This is the familiar pointing gesture. A *physiographic* depicts "either the form of a visual object or a spatial relationship (*iconographic* gesture), or that of a bodily action (*kinetographic* gesture). This is a gesture which concretely "acts out" an object or an action. One of Efron's many examples accompanies the words "so I finally wrote to him". Simultaneously, the speaker uses the index finger of one hand to write upon the other hand (p. 125). Finally, a *symbolic* or *emblematic* refers to the emblems I've discussed above, such as the "OK" sign in North America, which have conventionalized, agreed-upon meanings.

The gestures in Efron's other main category derive meaning in conjunction with speech. These he labels *logical-discursive*. Logical-discursive movements are "a kind of gestural portrayal, not of the object of reference … but of the course of the ideational process itself". "They are related more to the 'how' than to the 'what' of the ideas they reenact" (pp. 98-99). In other words, they portray the process of thought, rather than any object.

[2] As a courtesy to the reader, let me mention that in this literature review I'll be presenting a number of gesture classification schemes. While all are of interest, the one I'll be using in my dissertation will be the last one I present, McNeill's (1992). I'll compare the others to McNeill's in Section 2.1.2.11, when I summarize my literature review.

Two subcategories of this type are *batons* and *ideographics*. Batons are rhythmic gestures which Efron equates to a conductor's baton, used to "beat the tempo of … mental locomotion" (p. 98). An ideographic is a gesture which "traces or sketches out in the air the 'paths' and 'directions' of the thought-pattern." As an example, Efron describes an arm zig-zagging in the air between the locations of two thought propositions, before coming to a rest on one of the propositions, which is the speaker's conclusion (p. 99).

2.1.2.3 McQuown et al.

Beginning in 1955, and continuing through the 1960's, a group of researchers collaborated on analyzing a filmed interview between the anthropologist Bateson and a woman in her twenties (McQuown 1971). The collaborators included Bateson, the psychiatrists Fromm-Reichman and Brosin, the linguists McQuown and Hockett[3] (and later Trager and Smith), and the kinesicist Birdwhistell. The goal was multidisciplinary support of psychiatry. McQuown, for example, related prosody to behavior, noting that it could indicate normal, editorial, or introspective behavior, or that a narrowed pitch register could indicate depression, apathy, or boredom. There was no precise one-to-one temporal relationship found between speech and movement, perhaps in part because the researchers from each field looked at phenomena of different sizes. However, the group did recognize that speech and movement were fundamentally connected. Bateson noted, "It is, after all, only an historic accident … that linguists study data which can be heard, while the kinesicist studies data which can be seen. That the scientists have become specialized in this particular way does not indicate a fundamental separateness between the modalities in the stream of communication" (McQuown 1971, 19).

[3] Hockett participated in another well-known microanalysis of a psychiatric interview, "The First Five Minutes", by Pittenger, Hockett, and Danehy (1960). As the group had no kinesicist, they opted to analyze only the audio tape, and thereby forego looking at body movement. However, they did make the general observation that head and hand movements accompany speech stress.

At this point in the survey of gesture studies, it is appropriate to mention the intellectual debt to Erving Goffman for championing the study of face-to-face interaction as a discipline in its own right. He named this discipline the "interaction order—a domain whose preferred method of study is microanalysis" (Goffman 1983). Goffman inspired and was inspired by many of the researchers discussed in this section. He acknowledged the influence of Bateson, was taught as an undergraduate by Birdwhistell, and approved of the work by McQuown et al. as similar in spirit to his own (Kendon 1988a, pp. 20-21). Goffman's lifelong interest in what people "give off" when they are "co-present" with others gave direction to half a century of researchers to date. Kendon summarizes: "Goffman's [focus] was on how [interaction] was done; upon how, indeed, it was possible at all" (Kendon 1988a, p. 19).

2.1.2.4 Birdwhistell

Of the above researchers, Birdwhistell deserves particular mention. He has been regarded as the founder of the study of kinesics (Bateson, in McQuown 1971, p. 20), and was among the first to note a relationship between body motion and language. He regarded linguistic and kinesic markers as "alloforms, that is, structural variants of each other", and proclaimed that "linguists must turn to body motion for data to make sense out of a number of areas now hidden in the parts of speech" (Birdwhistell 1970, p. 127). Commenting on this quote, McClave observes, "If such a position proves true, resolution of some of the unresolved theoretical issues of linguistics may never be possible working within the traditional linguistic framework alone" (McClave 1991, p. 12).

Birdwhistell felt (1952, 1970) that gesture is structured in units similar to those in language. Analogous to phones, allophones, phonemes, and morphemes, he proposed kinesic counterparts termed *kines*, *allokines*, *kinemes*, and *kinemorphemes*. In terms of McNeill's dimensions of Kendon's continuum, Birdwhistell's view of gesture would be towards the right-

hand side – that is, it had segmented linguistic elements, which analytically composed meaning. However, Birdwhistell admitted that this theoretical goal was elusive:

> For several years I have been hopeful that systematic research would reveal a strict hierarchical development in which kines could be derived from articulations, kinemorphs from complexes of kines, and that kinemorphs would be assembled by a grammar into what might be regarded as a kinesic sentence.… I am forced to report that so far I have been unable to discover such a grammar. Neither have I been able to isolate the simple hierarchy which I sought. (McQuown 1971, chap. 3)

Birdwhistell also observed a tendency for kinesic stress to coincide with linguistic stress. Continuing his kinesic-linguistic analogy, he noted four levels of kinesic stress which in general corresponded to the then-hypothesized four levels of linguistic stress (1970, pp. 128-143). Beyond these observations however, he made no specific claims about the relationship of gesture to intonation, though he suspected a correlation. In his book *Kinesics and Context* (1970), he acknowledged: "Kenneth Pike … points out … [pitch] may contain some of the secrets of linguistic-kinesic independence …" (1970, p. xiv). Later in the book Birdwhistell noted, "It seems likely from preliminary data that some kind of systematic relationship exists between certain stretches of kinesic behavior and certain aspects of American English intonation behavior. However, the data are exceedingly elusive and must be investigated further before even tentative generalizations can be made" (1970, pp. 128-129). In spite of these thoughts, Birdwhistell did not pursue this line of investigation.

Birdwhistell collaborated closely with Albert Scheflen, who, like the McQuown group, studied body movement in support of psychiatry. Scheflen (1964, 1968) analyzed how body posture defined units within interactions. For instance, head shifts correlated with different points a speaker was making, and torso shifts with larger units of conversation. He noted how interactants often assumed similar postures, and if one interactant shifted posture, others followed suit. Scheflen and Birdwhistell also observed that eyeblinks, head nods, and hand movements

occur at intonational junctures (the end of a clause), and that the head and hand would rise with rising terminal pitch, and fall with falling terminal pitch (Scheflen 1964, pp. 320-321).

2.1.2.5 Pike

Pike (1967) felt that language and non-verbal activities were both elements of a larger, unified theory of human behavior. He extended the "-etic" and "-emic" concepts of linguistics into human behavior in general. A *behavioreme* was an *emic* (meaningful) element of some purposeful human activity (such as a church service, football game, or meal). The minimal unit of a behavioreme was an *acteme*, a term he credited to Zipf (1935). A verbal acteme was a phoneme, while a non-verbal acteme was Birdwhistell's kineme. Apart from his theoretical speculations regarding the relationship between speech and body movement, Pike undertook no analysis in the style of Birdwhistell.

2.1.2.6 Freedman and Hoffman

Freedman and Hoffman's (1967) motivation for studying body movement, like that of McQuown et al., was in support of psychiatry. They classified hand movements into two broad categories. In *body-focused* movements, which were not speech-related, the hands touch the body (e.g. grooming). Non-body-focused movements, or *object-focused* movements, are the speech-related movements of concern to the present study.

Object-focused movements are of five types, along a scale of increasing information content, increasing primacy vis-à-vis speech, and decreasing integration with speech.

Punctuating movements accentuate the accompanying speech and are well synchronized with it, but provide no additional information. They are seen as ancillary to speech.

Minor qualifying movements are a "simple turning of the hand from the wrists", which do add some information to the speech.

Literal-reproductive movements are an often complex "literal description with the hands" accompanying the speech. These portray a concrete object or event: for example, "a patient describing a man going down the stairs used her hand to indicate the descent through several steps" (p. 532). Literal-reproductive movements are equally "primary" with speech, and therefore less well integrated with speech.

Literal-concretization movements also describe an event, but not one with a physical referent (which is what literal-reproductive movements do). Literal-concretization movements describe something more abstract, for example an emotion. Since the event has no physical referent, this type of movement "concretizes" the abstract by making a metaphorical gesture about it. "For example, a patient, as he said he felt all mixed up, rotates his two hands in an attempt to concretize the feeling" (p. 532).

Finally, *major-qualifying movements* are poorly integrated with speech and are in fact disruptive to it. An example is vaguely groping with the hands. These are seen as primary over speech.

2.1.2.7 Ekman and Friesen

Ekman and Friesen (1969) refined Efron's taxonomy of gestures, and added some more. Three of their five categories do not fall within the term *gesture* as defined above. *Emblems* such as the "OK" sign in North America have conventionalized meaning. Also outside the scope of gesture are *adaptors* (renamed *manipulators* by Ekman [1999]). So-called *object adaptors* originate with movements for manipulating objects, *alter-adaptors* have to do with interpersonal contacts, and *self-adaptors* have to do with satisfying bodily needs, such as rubbing the eyes. A third type of non-gesture movement is *affect displays* (or *emotional expressions*, in Ekman [1999]) which display emotion. These happen mainly on the face, and the authors postulate that they may be universal.

The other two of Ekman and Friesen's categories would be considered gesture. The first is *regulators*, which "maintain and regulate the back-and-forth nature of speaking and listening between two or more interactants" (Ekman and Friesen 1969, p. 82). Examples include head nods, eye contacts, and postural shifts. The other gesture category is "illustrators", which are directly related to speech, and the sub-types of this category are what most gesture researchers focus on.

Of the six original types of illustrators, most are borrowed from Efron. *Batons* are used for emphasis. *Pictographs* resemble their referents. *Kinetographs* resemble a bodily action. *Ideographs* portray the course of thought. *Deictics* are the familiar pointing gestures, and *spatials* depict spatial relationships. Ekman (1999, p. 47) added a seventh type, *rhythmic movements*, which "depict the rhythm or pacing of an event".

2.1.2.8 Condon

Condon, in a series of works from the early 1960's carried out often in collaboration with W. Ogston, carried out detailed empirical analyses of the relationship of body movement and speech. Like McQuown et al., Condon's motivation was to support psychiatry. Condon refined the use of a hand-operated sound film projector, used in conjunction with a time-aligned oscilloscope, for frame-by-frame 'linguistic-kinesic microanalysis'. He observed that the parts of the body move in a hierarchical fashion, in which smaller, faster movements by some body parts (e.g. the hands or the eyebrows) are superimposed on larger, slower movements by other parts (e.g. the arms or the head). The points at which all these parts change direction or velocity coincide, so that the smaller movements are contained within the larger ones, like smaller waves within larger ones. Remarkably, he noticed that the smallest of these waves, very subtle changes in direction, correspond with the phones in the speech stream. These smallest waves Condon termed *process units* (Condon and Ogston 1966). He emphasized that the basic units of

movement are not discrete "things" glued together to make larger things (as Birdwhistell would have it). Rather, the units are seen as *patterns* in the constantly changing stream of *continuous* movement – hence the name *process* units.

Condon showed that the boundaries of the gestural waves coincide with the boundaries of units in the speech stream. For example, the more slowly changing movements of the head may align with larger spoken units, such as phrases, while the more rapid movements of the wrist and finger may align with syllables and even phones (Condon and Ogston 1967; Condon 1964 [via Kendon 1972]). Not only are these speech/gesture waves hierarchical, but they also occur rhythmically. In fact, Condon proposed a "rhythm hierarchy", with five levels (Condon 1976):

Phones
Syllable
Words
Half-second cycle[4]
One-second cycle

Both speech and gesture follow this rhythm hierarchy. For example, subtle body movements pattern at the level of the phone, changing direction and quality from one phone to the next. The same is true at the syllable and word level. Cycles of verbal stress and body movement tend to occur also at half-second and full-second intervals. This relationship of gesture and speech Condon termed *self-synchrony*, and he suspected it was due to a common neurological basis of both modalities. He believed that all behavior was fundamentally rhythmic.

In addition, Condon also made the "surprising and unexpected observation that listeners move in precise synchrony with the articulatory structure of the speaker's speech … In essence, the listener moves in synchrony with the speaker's speech almost as well as the speaker does" (1976, p. 305). Condon termed this *interactional synchrony*. "Metaphorically, it is as if the

[4] Condon's published data reveals that between two and three words typically occurred in a half second. These words might correspond to a stress foot, though Condon makes no mention of this.

listener's whole body were dancing in precise and fluid accompaniment to the speech" (1976, p. 306). He noticed it in infants as young as 20 minutes old. Just as he believed *self*-synchrony was based on a common neurological basis for speech and gesture, so Condon believed *interactional* synchrony was based on a common neurological basis for both speaking and listening. Noting that listeners react within 50 msec of a speaker's actions, he felt that the rapid response capability of humans (shown to be less than 10 msec) allows a listener to quickly "entrain" their own biological rhythm with that of the speaker.

Erickson (Erickson 1981; Erickson and Schultz 1982), in detailed audio-video analyses of human conversations, reported findings similar to Condon in terms of rhythm, self-synchrony, and interactional synchrony. Scollon (1981a, 1981b), building on Erickson's work, described how talk is timed to an underlying tempo, which binds conversational participants together.

2.1.2.9 Kendon

Kendon, starting from the late 1960's, built on Condon's work by more precisely developing the hierarchical units of gesture (and, incidentally, of speech, when he found little intonational work above the sentence level.) He used the same methodology as Condon (projector and oscilloscope) to microanalyze a 90-second discourse by a male speaker in a multi-party conversation. The gestural hierarchy he developed is as follows (Kendon 1972, 1980).

The smallest unit of motion is the *stroke* of the gesture, the short, dynamic "peak" of movement. The stroke may be preceded by an optional preparatory phase, in which the articulator (e.g. the hand) is brought to the point in space where the stroke occurs, and followed by an optional retraction phase, in which the articulator is brought either back to its starting point or to another point to begin the next gesture. This three-part combination of preparation, stroke, and retraction is called the *gestural phrase*. Gestural phrases may also have optional holds before and after the stroke, so the gestural stream can wait for the speech stream to catch up and stay

synchronized. Gestural phrases may flow from one to the other; a series of these, bounded by resting phases, is called a *gestural unit*. Gestural units are therefore the series of movements from one resting phase to another. Gestural units may be grouped together by a common feature, typically consistent head movement across the gestural units. Finally, at the highest level, Kendon noted a consistency of arm use and body posture.

Kendon observed a remarkable alignment of this gestural hierarchy with an intonational hierarchy, as shown in Table 1. (To Kendon's original table I've added my own "notes" column in Table 1). The gestural stroke typically occurred just prior to or exactly at the onset of a *stressed syllable*. The gestural phrase boundaries coincided with the edges of a *tone group*, a construct from the British School of intonation which is "the smallest grouping of syllables over which a completed intonation tune occurs" (Kendon 1972, p. 184). (See Section 2.2 below for McClave's working definition of tone group boundaries, based on Cruttenden's criteria). Tone groups have also been called *prosodic phrases, phonemic clauses, intonation groups*, and *tone units* in the literature. For consistency with the majority of researchers in this literature review, I will use *tone units* henceforth when discussing prior research.

The next level up, the gestural unit, coincided with what Kendon termed a *locution*, corresponding to a complete sentence. Groups of gestural units sharing consistent head movement were time-aligned with *locution groups*, or locutions sharing a common intonational feature apart from other groupings of locutions. For instance, the locutions within a locution group might all end with *low-rise*. Finally, consistent arm use and body posture were synchronized with Kendon's *locution cluster*, or paragraph. For example, Kendon's speaker used his left arm for gestures throughout his first paragraph, his right arm for the second, and both arms for his third and final paragraph.

Table 1

Kendon's Alignment of Gestural and Intonational Hierarchies

Gesture	Intonation	Notes
Stroke	Stressed Syllable	
Gestural Phrase (preparation + stroke + retraction)	Tone Unit ("completed intonation tune")	A "tone unit" would probably correspond to an intonational phrase
Gestural Unit (from rest to rest)	Locution (sentence)	
Consistent Head Movement	Locution Group (common intonational feature)	A "locution group" would consist of sentences with "parallel" intonation
Consistent Arm Use and Body Posture	Locution Cluster (paragraph)	A "locution cluster" would probably correspond to a "discourse segment"

Kendon also related head movement to sentence function, noting that a lowered head was found in sentences that moved the discourse forward. However, McClave (1991), in a point relevant to the present work, re-analyzed Kendon's clearly-transcribed data to show a potential correlation between head movement and intonation, noting "The general pattern emerges that the head lowers with falling pitch and rises with rising pitch" (McClave 1991, p. 31).

Kendon is credited with the insight that speech and gesture are not just somehow connected, but actually are two surface forms of a single underlying utterance. He in particular challenged the notion that gesture was an after-occurring by-product of speech. The tight synchrony between the two modalities, and the fact that strokes occur just *before* stressed syllables, led Kendon to the conclusion that gesture did not arise *from* speech, but rather that the two had a common origin. He terms this origin an *idea unit*. In a similar vein, he later (1988b, p. 264) equated tone units (and thus their accompanying gestural phrases) more or less with Halliday's (1985) *information units*, or minimal units of sense.

Kita et al. (1997) proposed refinements to parts of Kendon's taxonomy. Most generally, they introduced the terms *movement phase* and *movement unit*, instead of *gesture phase* and *gesture unit*, to capture generalities between gesture and sign language movements. They differentiated between a *dependent hold* - parasitic to a stroke - and an *independent hold* – an expressive phase in its own right. Kita et al. also noted that preparations optionally begin with a *liberating movement*, in which the hand frees itself from a resting position, and may also contain a *location preparation*, in which the hand changes location, and a *hand-internal preparation*, in which the takes on the form and orientation of the following stroke. Finally, the retraction is said to end at the first contact with the resting surface, excluding any settling-down motions.

2.1.2.10 McNeill

McNeill (in numerous works) built in turn upon Kendon's research. I will first present McNeill's taxonomy of gestures, and then discuss his views on the relationship of gesture with speech. As I will use McNeill's gesture classification in my own work, and as my hypotheses will be grounded in his theoretical viewpoint that speech and gesture spring from a common origin, I will review his work in more depth than the others in this section.

McNeill's classification (borrowing from Efron and others) includes four types (McNeill 1992).

Deictic gestures are the familiar pointing motions to identify an entity under discussion. A variant, *abstract deictics*, refer to locations in space where entities under discussion have been placed (McNeill et al. 1993).

Iconic gestures represent a concrete idea. For example, a speaker retelling a scene from a Sylvester and Tweety Bird cartoon, in which Tweety Bird stuffs a bowling ball down a drainpipe on top of Sylvester, may stuff one hand with fingers together inside a ring formed by the other hand.

Metaphoric gestures represent an abstract idea. An example is the concept of a story being presented as a *conduit* for the story line itself. In McNeill's example, a speaker informs his listener at the outset that "it was a Sylvester and Tweety Bird cartoon", while holding his hands in front of his torso about a foot apart, palms facing each other, as if holding a package. This package represents the story about to be presented.

The fourth type of gesture McNeill terms *beats*. This is another name for what others have termed *batons*, which are timed with the "rhythm" of speech. Beats vary in size, and can be large, noticeable movements. Often, however, they are small, barely perceptible flicks of the wrist or finger, occurring on stressed syllables (though the converse is not true; not all stressed syllables are accompanied by beats). A distinguishing feature of beats is that they have only two phases (typically a down-up pattern), as opposed to the other types, which often have three (preparation-stroke-retraction).

As an aside, Kendon (1983) noted that beats might often retain the same handshape form of a preceding iconic within the same gestural unit. Tuite (1993) ascribes a cohesive function to these "inertial" or "anaphoric" beats, linking their lexical counterparts back to earlier references that occurred with the same handshape.

Another label, *representational gestures*, has also been used in the literature to (usually) refer to iconics, metaphorics, and sometimes abstract deictics.

While deictics, iconics, and metaphorics carry propositional meaning, beats are non-propositional. They have instead a discourse function, being "extranarrative". In the realm of narrative, for example, speakers can operate on one of at least three levels. The first is "narrative", in which the speaker is relaying simple, chronological events of the story line itself. The second is "metanarrative", in which the speaker explicitly conveys information about the structure of the narrative, by talking "about" the story. Examples may include the introduction to

the story, or backtracking to fill in previous events. Finally, in the "paranarrative" role, the speaker steps out of the storytelling role altogether, and interacts as fellow conversant with her listeners. One contribution of beat gestures, according to McNeill, is to signal that the speaker is shifting from one of these narrative levels to another. The utterance doing the shifting may contain several beats, sprinkled throughout its stressed syllables.

Having discussed McNeill's typology of gestures, I'll now turn to his views on the relationship between gesture and speech. In his 1985 article "So You Think Gestures are Nonverbal?", McNeill explicitly states that gesture is in fact "verbal", and is linked with speech such that the two modalities "are part of the same psychological structure and share a computational stage" (1985, p. 350). He gives five arguments as evidence.

First, gestures occur *only* during speech. Gestures as defined above do not occur in non-speech situations, and occur only extremely rarely in listeners. More specifically, the majority of gestures (90% by McNeill's count) occur during the speaker's actual articulation, and not, for example, during pauses.

Second, gestures and speech have parallel semantic and pragmatic functions. All four types of gestures convey related content (or perform related discourse functions) to their lexical counterparts. Often the content of gesture complements that of speech, by providing, for example, additional information such as a manner of action, or the physical relationship of two entities.

Third, speech and gesture are temporally synchronized. Gestures not only convey content parallel to their lexical counterpart, but they do so at the same time. In fact, speakers seem to adjust the timing of their gestures, performing a hold before or after the stroke, to ensure this synchronization. And gestures "almost never cross clause boundaries" (1985, pp. 360-361), ensuring that they stay within their lexical counterpart's propositional phrase.

Fourth, gestures and speech are affected in parallel ways by aphasia. Broca's aphasics speak "telegraphically", with command of content words but not of relating these words into fluent sentences. Their gestures consist of numerous iconics (paralleling the content words) but few beats (paralleling the lack of higher-level discourse ability). In contrast, Wernicke's aphasics speak "vacuously", with fluent sentences but little concrete semantics. Their gestures contain beats and occasional metaphorics, but few or no iconics.

Finally, the development of gesture parallels that of speech in children. Speech abilities progress roughly from deictics and concrete words to grammatical relations to discourse coding. Gesture abilities progress from deictics and iconics to metaphorics and beats.

As an aside, Butcher (1994) showed that temporal and semantic linkage between the two modalities does not develop in children until the age of around a year and a half.

Based on the above, McNeill proposes three "synchrony rules" (1992, pp. 26-29). Two of them, the *semantic synchrony rule* and the *pragmatic synchrony rule*, simply state that if speech and gesture co-occur they must present the same semantic information or perform the same pragmatic function. The third is more relevant to my dissertation:

Phonological Synchrony Rule: "… the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech (Kendon 1980)" (McNeill 1992, p. 26).

As mentioned, McNeill agreed with Kendon that gesture and speech share a common origin. He called this (theoretical) origin the *growth point*, from which the utterance (in all its modalities) is "unpacked" (McNeill 1992, 1997, 1999, 2000b; McNeill and Susan Duncan[5] 2000). The growth point is born as the "novel departure of thought from the presupposed background" (McNeill 1992, p. 220). It incorporates context by definition, as it marks the "newsworthy"

[5] As Susan D. Duncan and Starkey D. Duncan have the same initials, I will include their first names when referring to them, to avoid confusion.

contrast to the immediate context of the discourse (McNeill 2000b). The growth point contains properties of both gesture and speech: imagistic *and* linguistic, global *and* segmented. The imagistic part becomes the gesture, the linguistic part the speech. The two modalities continually interact during the "unpacking" to reach their surface forms. The contents of the growth point can be inferred from these surface forms combined, based on the timing of the gesture. Because the gesture is less encumbered by linguistic processing, and because it thus can surface sooner (explaining its slight precedence to speech), the growth point is revealed in the speech and gesture on or after the gestural stroke, but never before.

McNeill has furthered Kendon's analyses in another way. As discussed above, Kendon (1972) noted that speakers tended to use similar gestures throughout a single locution cluster (or paragraph). McNeill gave the term *catchment* to this "recurrence of gesture features over a stretch of discourse. … Catchments are recognized from two or more gestures (not necessarily consecutive) with partially or fully recurring features of shape, movement, space, orientation, dynamics, etc. (McNeill 2000b, p. 316)". Kendon also noted (as discussed above: see Table 1) an integration of gesture, intonation, and discourse structure. McNeill furthers Kendon's analysis of the aligned hierarchies, and proposes catchments as "the locus around which this integration proceeds" (McNeill et al. 2001b, p. 1).

To illustrate this point, McNeill, Quek, and colleagues (McNeill et al. 2001a, 2001b, Quek et al. 2000, 2001) first segmented a narrative (of a subject describing her house, or the layout of a town) into sections based on catchments. They then independently segmented the text of the narrative into a hierarchical discourse structure, based on guidelines published by Grosz and colleagues (Nakatani 1995). They found nearly perfect alignment between gestural catchments and discourse structure, "implying that the gesture features were not generated

haphazardly, but arose as part of a structured, multi-level process of discourse-building." (Quek et al. 2001, p. 19).

One final observation of McNeill's is worth discussing, that of rhythm. McNeill's colleague Tuite (1993, p. 99) postulated an underlying "rhythmic pulse" influencing both speech and gesture. The surface correlates of the rhythm are the stroke (for gesture) and the tonic nucleus (for speech), which are correlated such that the stroke occurs just before or at the nucleus. McNeill measured this rhythm in examples where speakers gestured fairly continuously, finding that the period between strokes was fairly regular, and ranged between one and two seconds, depending on the speaker (McNeill 1992, p. 244).

2.1.2.11 Summary of Earlier Gesture Studies

I have discussed a variety of gesture researchers in the preceding sections. I'll briefly recap them here, with an eye towards leaving the reader with the important points relevant to my dissertation. Of the researchers I've touched on, four have proposed classifications of gesture types—I'll also summarize and compare those at the end of this section.

I first defined gesture in terms of Kendon's and McNeill's continuum, in which gesture is the opposite of sign language. Gestural movements have no predefined meaning, but rather are interpreted in conjunction with obligatory accompanying speech.

I then surveyed several millennia of classical studies of gesture, ranging from prescription of proper rhetorical gestures, to musings on the meanings and origins of gestures.

In the modern era, Efron was the first to carefully study gestures, and proposed a classification scheme. In the mid-20th century, a number of psychologists studied body movement in relation to speech, as chronicled by McQuown. Their collaborator Birdwhistell pioneered the study of kinesics, and credited Pike with looking to pitch as a part of a linguistic-kinesic interdependence.

Two pairs of researchers—Freedman and Hoffman, and Ekman and Friesen—each proposed a gesture taxonomy, the latter pair starting with Efron's work. Condon pioneered micro-analysis using a sound projector and oscilloscope, and discovered fine-grained relationships not only between body and speech, but also between speakers. Kendon advanced Condon's work, and discovered a hierarchical relation between gesture and intonation. Finally, McNeill, while refining earlier taxonomies, argued clearly and persuasively that gesture and speech share a common cognitive origin.

Along the way I described four different classifications: Efron's, Freedman and Hoffman's, Ekman and Friesen's, and McNeill's[6]. For the convenience of the reader, I will reproduce part of McNeill's table correlating his gesture categories with those of Efron and with the "illustrators" (speech-related gestures) of Ekman/Friesen and the "object-focused" (speech-related) movements of Freedman/Hoffman. I'll discuss Table 2 in light of McNeill's terminology, as these are the terms I'll be using throughout the rest of my dissertation.

Table 2

Correlation of McNeill's, Efron's, Ekman/Friesen's, and Freedman/Hoffman's Gestures (adapted from McNeill 1992, p. 76)

McNeill	Efron	Ekman and Friesen	Freedman and Hoffman	Brief Definition
Iconics	Physiographics Kinetographics	Pictographs Kinetographs	Literal-Reproductive	Portray concrete idea
Metaphorics	Ideographics	Ideographs	Literal-Concretization	Portray abstract idea
Deictics	Deictics	Deictics		Pointing
Beats	Batons	Batons	Punctuating	Rhythmic movement

[6] Wundt (1921/1973) also had a classification scheme similar in spirit to these others. However, as his was mainly based on emblematic gestural "languages" such as those used by Plains Indians, Neopolitans, and Cistercian monks, I will not consider it here. His categories (e.g. *imitative*, *mimed*, *connotative*, *symbolic*) had to do with the level of physical similarity or abstraction between the emblem and its meaning.

McNeill's *iconic* gestures concretely portray the idea being expressed. This sense of concrete portrayal is captured in the names the other researchers have given to this type of gesture; e.g. *physiographic*, *kinetographic*, *pictograph*, and *literal-reproductive*.

McNeill's *metaphorics* portray a more abstract idea than a concrete object or action. An example is a story itself being presented, with a metaphoric gesture of presenting a package. This correlates with Efron's *ideographics* and Ekman and Friesen's *ideographs*, which portray a thought process. It also correlates with Freedman and Hoffman's *literal-concretization* movements, which "make concrete" an abstract idea, such as a package gesture for presenting a story.

Deictic gestures refer to pointing, and are identically named by all researchers who included them.

Finally, McNeill's *beats* are the same as Efron/Ekman and Friesen's *batons*, and Freedman and Hoffman's *punctuating movements*. As mentioned, beats vary in size. Their distinguishing characteristics are that they mark the tempo of speech, and have only two phases (typically down-up).

These, then, are the four categories I'll be classifying gestures into in my dissertation: iconics (concrete ideas), metaphorics (abstract ideas), deictics (pointing), and beats (rhythmic movements).

2.1.3 Functions (and Cognitive Origins) of Gesture

The above section discussed, among other things, the forms of gestures. These forms are fairly uncontroversial, in spite of the differing terminology. Such is not the case for the hypothesized *functions* of gesture, spurring a vigorous debate in the literature. I'll organize my review of research since McNeill (1992) along the lines of this debate. As the arguments in this debate relate to the purported cognitive origins of gesture, I'll discuss these also.

On one side of the debate, gesture is seen as communicative. It is generated for the benefit of the listener. It provides meaning apart from that provided by speech. In addition, although gesture and speech interact in their production of surface forms, neither is derivative from the other. Rather, they stem from a common source, and are produced in parallel. McNeill is a representative of this viewpoint, though there are many others.

On the other side of the debate, gesture is seen as assisting the production of speech. It is generated for the benefit of the speaker. It provides little or no meaning apart from that provided by speech. The production of gesture is thus derivative from the production of speech. A representative of this viewpoint is Butterworth (Butterworth and Beattie 1978; Butterworth and Hadar 1989; Hadar and Butterworth 1997), although again there are many others.

2.1.3.1 Is Gesture for Assisting Speech Production?

I will first present the arguments for the claim that gesture assists the speaker in the production of speech. Before I begin, I'll note that several researchers in the literature (e.g. (Krauss and Hadar 1999), have worded the claim by saying that gesture *facilitates* speech production. While this is completely appropriate terminology, its shorter version, that gesture is merely *facilitative*, can be construed two ways. Gesture can facilitate things for the speaker (by making speech production easier). Or, gesture can facilitate things for the listener (by being communicative), which is the opposite claim. The appropriate thing to say might be that gesture is *speech-production-facilitative*, but this is rather cumbersome. I'll therefore use the shorter (but perhaps inelegant) term *production-aiding*, when referring to the theory in general.

The production-aiding claim is partly based on observations by Goldman-Eisler and colleagues (Henderson et al. 1966; Goldman-Eisler 1967) that spontaneous speech is characterized by "alternating patterns of relatively *hesitant phases* with *high pause/speech ratios* and relatively *fluent phases* with *low pause/speech ratios*" (Nobe 1996, p. 7, emphasis in

original). That is, a speaker will have a "temporal cycle" consisting of a phase of relative silence with short vocalizations, followed by a phase of relatively fluent speech, with few hesitations. It was claimed that the hesitant phases were used for planning what the speaker was going to say, and the fluent phases for actually saying it. The fluent phase typically included several clauses. Thus, the overall cycle was called a "cognitive rhythm". Note that this is different than metrical rhythm, and had a much larger scale (covering multiple clauses, and a period of hesitations).

Butterworth and Beattie (1978) observed that representational gestures tended to have their onsets during *pauses* in the fluent (i.e. production) phases of the temporal cycle, and thus preceded their lexical counterparts. The reason for this, they hypothesized, was that difficulty in retrieving a lexical item would cause a hesitation. The gesture would then initiate to assist the lexical search, in a sort of "cross-modal priming" (Morrel-Samuels and Krauss 1992, p. 622). As evidence, they noted that more gestures were associated with low-frequency words, which presumably were more difficult to retrieve. In Butterworth and Hadar (1989), the authors postulate two origins for gesture, based on the amount of temporal asynchrony between the gesture and the associated word. A small asynchrony would mean the gesture originated in the lexical-search stage, while a large asynchrony would place the gesture's origin at some pre-verbal stage. Hadar et al. (1998) conducted brain lesion studies, in which patients with semantic and phonological impairments produced more gestures than a control group, again offered as support for the lexical-retrieval hypothesis of gesture.

Schegloff (1984) also noted that both the onset and "thrust" (stroke) of certain gestures preceded their lexical affiliate. His functional explanation was that gesture foreshadowed the speech. The occurrence of the speech then completed the intended meaning. Gesture was thus subservient to speech, used to "till the soil into which the words are dropped" (1984, p. 291). But this "tilling the soil" did not provide meaning apart from speech. "We regularly get their

[gestures'] sense … only when the bit of talk they were built to accompany arrives." (1984, p. 291).

Dittmann and Llewellyn (1969) also made use of the finding that movements tended to occur during or immediately after pauses. They hypothesized that difficulty in formulating speech (evidenced by pauses) results in tension, which builds up and "spills over" into motor activity. The motor activity may also be a signal to the listener that the speaker's concept is difficult to conceptualize.

Rimé (1982) provided empirical support for the common observation that people gesture when no listener is visible, supporting the view that gesture is for the benefit of the speaker. But then, he asked, why do people gesture? Because motor activity supports cognitive processing, he claimed, in a "cognitive-motor" theory of nonverbal behavior (Rimé 1983). Rimé and Schiaratura (1991), in contrast to the claim that gesture provides meaning above and beyond that of speech, claimed that gesture provides mostly information redundant to speech. In addition, they cited experiments showing that an increase in gesturing is actually detrimental to understanding the speech stream, stating that if a speaker "wants to be clearly understood, he or she should display as few gestures as possible" (1991, p. 276).

Aboudan and Beattie (1996), in experiments with an Arabic speaker, found that a lower pause-to-speech ratio during the planning phase (that is, the planning phase had fewer pauses than usual) resulted in more iconics during the fluent phase. Presumably, the less time spent planning, more difficulty retrieving lexical items during the production phase, and therefore more gestures required to assist the process.

Feyereisen (1997) carried out stimulus-response experiments similar to those of Levelt et al. (1985), testing deictic words and deictic gestures alone and in combination. His results (and Levelt et al.'s) implied that there was competition for processing resources between the two

modalities. This would explain the lexical-retrieval hypothesis' reason for the temporal asynchrony. In an overview of neuropsychological evidence, Feyereisen (1999) noted little support for a "common cerebral basis for speech and gestures" (1999, p. 20). He based this on the existence of differing subsystems (not single systems) of both speech and gesture production, which break down in diverse ways in aphasia and limb apraxia, respectively.

Krauss and colleagues (Krauss et al. 1991) empirically tested whether hand gestures convey information beyond that of speech. They found this true to only a limited extent, and that the information conveyed was largely redundant with speech. They also reported (Morrel-Samuels and Krauss 1992) that the less familiar a lexical item, the more the lexical item was preceded by gesture. They acknowledged that gesture may not necessarily facilitate lexical retrieval (though this is "plausible"), but that difficulty in lexical retrieval could certainly account for the greater delay. Later (Krauss and Hadar 1999, p. 105), they strengthened their claim, stating simply that "lexical gestures facilitate lexical retrieval". (Their term *lexical gestures* refers to co-speech gestures with the exclusion of deictics and beats.)

Krauss and colleagues also proposed not only a mechanism for facilitation of lexical retrieval, but also a model for the production of lexical gestures (Krauss and Hadar 1999; Krauss et al. 2000). Their model is an extension of Levelt's (1989) model of speaking which includes three phases: (1) *conceptualizing* a pre-verbal message, (2) *formulating* a phonetic plan (via syntactic/morphologic/phonological encoding), and (3) *articulating* speech. Krauss and his colleagues added a parallel motor pathway, starting with spatial/dynamic features of the source concept, and continuing through a motor planner to a gesture. A kinesic monitor of this gesture then fed the spatial/dynamic features of the source concept back into the speech model, at the level of the phonological encoder. Because the phonological encoder is also involved in lexical retrieval, the gesture thus facilitated lexical retrieval, via cross-modal priming. Krauss and

colleagues state that any information contained in the gesture is *not necessarily* intended to be communicative to the listener. Rather (in my understanding), gesture is a carrier of certain features to aid the lexical retrieval module, and this carrier just happens to be a visible one.

2.1.3.2 Is Gesture for Communicating?

I will now present the arguments in the literature for the claim that gesture is communicative, is for the benefit of the listener, provides meaning apart from that of speech, stems from a common source with speech, and is produced in an interactive, parallel fashion.

As mentioned, McNeill is perhaps the most visible proponent of this view. As I have already discussed his arguments, I will not repeat them here.

Kendon (1993), re-emphasizing the temporal synchronization between the two modalities, provided examples of both gestures waiting for speech to catch up, and vice versa. (McClave (1991) found similar examples.) Kendon commented that "Examples such as these show that the movements of gesticulation … and speech together are deployed in a unitary program of execution, adjusting one to the other" (1993, p. 46).

Nobe (1996, 2000) re-analyzed McNeill's data, specifically to test whether gestures tended to have their onsets in the *pauses* of fluent phases (as Butterworth and Beattie (1978) claimed), or whether they had onsets during actual phonation. He confirmed McNeill's analysis (but see Beattie and Aboudan's (1994) comments below on the effect of social context on McNeill's data).

Susan Duncan (1996) studied how gesture relates to "thinking-for-speaking", a term she credits to Slobin (1987) that describes the cognitive process of preparing an utterance. Duncan, a collaborator of McNeill, is interested in how the growth point operates. Duncan looked at speakers in two typologically different languages—Mandarin Chinese and English—and the gestures they made co-occurring with two linguistic structures that differ widely between the two

languages—verb aspect and topic prominence. She found that speakers of both languages produced similar types of gestures denoting verb aspect, suggesting that the languages' surface differences were somewhat superficial. In contrast, the gestures related to topic prominence differed between the two populations, suggesting that topic prominence may operate at a more fundamental level in thinking-for-speaking. In either event, the point relevant to the debate under discussion is that gesture communicates information apart from that carried in speech—in this case, verb aspect and topic prominence.

Beattie, an early proponent of the production-aiding view, has recently put forth evidence for the communicative theory. Beattie and Coughlan (1999) asked subjects to repeat cartoon narratives several times in succession, presumably accessing lexical items more readily with each repetition. Contrary to the lexical-retrieval hypothesis, the number of iconic gestures did not diminish. The same study also employed the tip-of-the-tongue (TOT) state, in which speakers are momentarily unable to retrieve a word. When subjects were not allowed to gesture, TOT resolution did not become more difficult, as the lexical-retrieval hypothesis would predict. In contrast, TOT resolution became easier.

Beattie and Shovelton (1999) tested McNeill's claim that gestures convey information beyond that of speech. They played to subjects clips of narration (containing both speech and gesture) in audio only, video only, and both. Then they used structured questionnaires to determine what the subjects had absorbed. They found that gestures did convey additional information beyond that of speech, but that this information was limited to the size and relative position of objects in the narration. In a similar experiment (Beattie and Shovelton 2003), they produced two versions of a television ad, which differed only in the gestures accompanying the identical speech track. Again, more information uptake took place among viewers when gestures were added.

Kelly and Barr (1999) also tested McNeill's claim about the relative information in gesture. Subjects watched videos of scenes portrayed by professional actors who were instructed by the experimenters to use certain gestures. The gestures were designed to vary by information content. By structured interviews, the experimenters found that subjects did incorporate the information from gestures. Moreover, the subjects were unable (when explicitly asked) to say in which medium they had received the information, lending support to a common psychological base for both modalities. Cassell et al. (1999; McNeill 1992) performed a similar experiment, with an experimenter intentionally mismatching the content of the gesture with that of the speech. The subjects nonetheless incorporated the gestural information.

Bavelas and colleagues (Bavelas et al. 1992; Bavelas and Chovil 1995) carried out experiments showing that more gestures occur in dialogues than monologues. Bavelas et al. argue that this increase in gesture is proof that gestures are communicative. They then proposed re-dividing the illustrators (speech-accompanying gestures) into two classes. *Topic* gestures refer to the topic. *Interactive* gestures refer instead to the interlocutor, and "help maintain the conversation as a social system" (Bavelas et al. 1992, p. 469). Interactive gestures include beats, but also some non-beats, such as deictics and metaphorics (e.g. a conduit metaphor passes something to the interlocutor). Interactive gestures are themselves subdivided into major and minor categories. Major categories include *delivery* gestures (e.g. marking information status as new, shared, digression), *citing* gestures (acknowledging others' prior contributions), *seeking* gestures (seeking agreement, or help in finding a word), and *turn coordination* gestures (e.g. taking or giving the turn).

Thus, by Bavelas' definition, at least some gestures (the interactive ones) are communicative, and would therefore support the pro-communicative side of the debate. However, this communication is limited to interactional purposes, and explicitly (by definition)

excludes communication of content apart from speech. Bavelas' interactive gestures would therefore not in themselves support a strong version of the communicative theory of gesture, which says that gestures communicate content apart from speech. But they do support the communicative stance in general. Gill et al. (1999) noted similar functions of gesture, adding body movements to the repertoire of pragmatic acts used in dialogue act theory (e.g. turn-taking, grounding, acknowledgements).

McClave (1991) provided clear empirical evidence that if more than one gesture occurs in a tone unit, the gestures are compressed and "fronted" to all finish before a stressed syllable. Like Kendon's discovery that gestural strokes precede their counterpart words, McClave's gestural-fronting evidence supports the part of the communicative hypothesis which claims that gesture does not arise *from* speech, but rather that the two share a common origin. McClave will be discussed at more length in the review of intonation literature below.

Clark (1996) viewed gestures as clearly communicative, placing them, along with words, in what he terms *primary* (communicative) and *secondary* (meta-communicative) conversational "tracks". He rejected the argument of Krauss et al. (1991) that, because many iconic gestures are uninterpretable without accompanying speech, they could not be communicative. Clark pointed out that "Most words aren't fully interpretable when isolated from their spoken contexts, yet words are patently communicative. Gestural utterances are no different" (1996, p. 179).

Emmorey (1999) investigated whether native (non-speaking) signers of sign language produced gestures akin to those accompanying spoken language. While signers did not make manual gestures concurrent with signing, they did alternate signing with non-sign, non-linguistic gestures. Emmorey doubted that these gestures facilitated lexical access, as they were not associated with particular lexical items, as spoken gestures are. Rather, they appeared to be communicative, depicting an event being described by the signer.

Mayberry and Jaques (2000) examined gesturing during stuttered speech. They found that gestural and speech execution remained tightly coupled temporally throughout the disruptive bouts of stuttering. Gesturing hands would either rest or freeze during the stuttering, immediately resuming the gesture upon the resumption of fluent speech. Furthermore, gesturing overall decreased for stutterers, parallelling the decrease in speech content. This supports the communicative theory, which says that speech and gesture are produced in parallel towards a common communicative goal, and counters the production-aiding theory, which predicts that gestures would be more frequent during stuttering to compensate for verbal disfluencies.

Mayberry and Jaques further speculated on the nature of speech-gesture integration. They subscribed to the "Dynamic Pattern" theory of Kelso et al. (1983), which states that "the coordination of movement of different limbs is thought to arise from interactions between oscillatory processes themselves rather than from central representations" (Mayberry and Jaques 2000, p. 210). Thus, a cycle of speech (based on prosodic patterns) and a cycle of gesture (based on stress patterns of gestural movement) are able to synchronize "on-line" during message execution, without control of a central representation. The authors suggest that these harmonizing oscillatory cycles may account for Tuite's "rhythmic pulse" underlying speech and gesture.

De Ruiter (2000) followed Krauss and colleagues above, in adapting Levelt's speech production model to gesture. Unlike Krauss and colleagues, however, de Ruiter followed McNeill in assuming that gesture and speech share a common origin, are planned by the same process, and are both part of the speaker's communicative intent. De Ruiter went beyond McNeill in trying to specify exactly how the common origin develops into surface speech and gesture, via a "Sketch Model" (so called because the conceptualization process outputs not only a pre-verbal message for speech formulation, but also a "sketch" of a gesture to a gesture planner).

In the Sketch Model, speech and gesture production are not continuously interacting, but there are specific signals exchanged to temporally synchronize the two modalities.

2.1.3.3 Studies Not Easily Classified in Either Side of the Debate

Some research can not easily be classified in either side of the debate, either because it remains agnostic, or because it embraces both sides.

Beattie and Aboudan (1994) is an example of the former. The authors noted that part of the overall argument stems from disagreement over whether or not gestures tended to have their onsets in pauses (McNeill and colleagues said no, Butterworth and colleagues said yes). Beattie and Aboudan speculated that the different kinds of data being examined might be the problem. In particular, they hypothesized that data from a two-way social interaction (as recorded by Butterworth and Beattie 1978) would contain more gestures in pauses than data from a more one-way monologue (as recorded by McNeill). The reason is that in a social situation, speakers may view silent pauses as opportunities for the listener to break in and take the floor. To prevent this, speakers may use gestures as "[interruption-]attempt-suppressing signals", as suggested by Starkey Duncan (Duncan 1972; Duncan and Fiske 1977). By varying the social setting and examining the number of gestures in pauses, Beattie and Aboudan confirmed their hypothesis that more social interaction resulted in more gestures in pauses. Reproducing these effects with an Arabic speaker, they then proposed that the effect of social context is universal (Aboudan and Beattie 1996).

Cohen (1977) empirically tested whether people gesture more in face-to-face situations than over an intercom, supporting the communicative function (in his words, gesture is for the "decoder"). He found this was indeed the case. He also independently tested whether people gesture more (per unit time) when giving complicated messages than when giving simple ones,

supporting the production-aiding function (gesture is for the "encoder"). He found this was also the case. Both functions of gesture were thereby shown to exist.

The fact that people gesture at all while speaking on the telephone (even though less than when face-to-face) has been given as evidence that gestures are for the benefit of the speaker, not the listener. A number of researchers have suggested, however, that people gesture on the phone simply out of habit. Bavelas (personal communication, November 3, 2003) tested not only the feature of *visibility* (face-to-face vs. telephone), but also the feature of *dialogue* (giving descriptions on the telephone vs. to a tape recorder). She found a decrease in the rate of gestures from face-to-face to telephone to tape recorder, indicating that people gesture on the phone partly because they know they are in dialogue with someone else. Bavelas also found that the qualitative feature of gestures changed as function of visibility. Face-to-face speakers produced gestures that were larger and not redundant with speech, while speakers in the telephone and tape-recorder conditions produced smaller, more redundant gestures. Gesturing on the telephone has therefore been claimed as evidence for both sides of the debate.

In the same study, Bavelas et al. also found possible evidence for the production-aiding hypothesis. They noted that "in the telephone and tape-recorder conditions, speakers would point at and trace their fingers over the picture they were describing, which rarely happened in the face-to-face condition; these may be 'self-prompting' gestures."

Goldin-Meadow and colleagues, in a series of works (Goldin-Meadow 1997, 2000; Goldin-Meadow and Sandhofer 1999) also described naturally-occurring "mismatches" between speech and gesture content, describing the gestural component as a window (for the listener) into the speaker's thoughts. While this may have a communicative effect, Goldin-Meadow and colleagues speculated that the gesture may assist the speaker's cognitive processing, by allowing the speaker to explore ideas that may be more difficult to process verbally.

Tuite (1993) specifically stated that gesture is production-aiding, yet his arguments are puzzlingly close to the communicative reasoning of his colleague McNeill. Tuite wrote: "*...the activity of gesture primarily occurs for the 'benefit' of the speaker/gesturer, and not for the listener*" (p. 94, italics in original). Following Cosnier (1982), Tuite explained that "gestural activity in some way assists the speaker in the process of encoding an underlying representation into the verbal modality" (p. 102). This assistance, then, is not lexical retrieval, but rather transferring meaning from McNeill's growth point to the surface speech. Gesture can do this because the imagistic origins of growth point are "not only visual – they may also include auditory, tactile, and kinesthetic elements. The body is … involved in the formation of concepts." Although the verbal component of the growth point must surface in a "rote-memorized lexical item, its imagistic counterpart often remains alive for speakers, as shown by gesture. Just as the body is intimately involved in the activity of perceiving and representing the world, so it is involved in the process of acting upon these representations in order to *communicate*" (p. 102, emphasis mine). Gesture, then, is an "epiphenomenon from a strictly communication-theoretic point of view", although a "necessary a component of the process." (p. 102).

To me, Tuite's argument that gesture *assists* communication by showing an imagistic or kinesthetic counterpart to speech, is tantalizingly close to saying gesture is communicative itself. For if this imagistic, kinesthetic element cannot be expressed by speech alone, is it not communicated via gesture? Tuite appears to follow McNeill's progression of speech and gesture from growth to surface form, yet denies that the gestural part is communicative.

In a related vein, a primary claim of McNeill's, throughout his works, is that "gestures … *help constitute thought*" (McNeill 1992, p. 245, italics in original). McNeill and Susan Duncan (2000, pp. 156-157) clearly state that gesture not only communicates cognition, but can be

cognition itself – "thought in action". They write, "… by gesture … cognitive being itself is changed." This shaping of one's own thought via gesture could be said to be production-aiding for the speaker, akin to thinking aloud.

Therefore, whereas Tuite appears to say that gestural thought assists communication (but is primarily production-aiding), McNeill and Duncan appear to say that gestural communication assists thought-production (but is primarily communicative). The two ideas are, of course, closely related, and there may only be a difference of perspective between them.

Kita (2000) explicitly stated that gesture is both communicative and production-aiding. To explain the evidence for the production-aiding view (e.g. speakers still gesture when no listener is visible), Kita put forth the "Information Packaging Hypothesis". He stated that certain gestures help speakers organize information into manageable packages suitable for expression in a single utterance or clause. Gestures do this because they allow access to a different way of thinking - spatio-motoric – in addition to the default analytic thinking of spoken language. Access to both modes of thinking, and the possibility of expression in both surface modalities, helps the speaker more efficiently organize and present information within the utterance.

Iverson and Goldin-Meadow (1999; Iverson et al. 2000) reported that children blind since birth definitely gesture, although not always with comparable form, content, and quantity as their sighted peers. The researchers agreed with McNeill that gesture plays a communicative role. Yet they reasoned that because blind children have never seen gestures or experienced firsthand their communicative value, the fact that they gesture may mean that gesture plays a role for the speaker as well as for the listener. This role may be to help the children think through the problem. Thus, Iverson and Goldin-Meadow believed, like Kita, that gesture may be both communicative and production-aiding.

Goodwin and Goodwin (1986) analyzed the head, face, and hand movements of speakers during word searches. They speculated that because the head and face movements are so stereotypic (averted gaze and a "thinking face"), they may not be communicative, and may be "simply adjustments to the cognitive demands that a word search imposes" (1986, p. 58). Yet a main point of their article is that listeners can pick up on speaker's movements during word searches, and in response actively suggest the missing word. In one example, "The [hand] gesture occurs at the moment where a change in coparticipation status is occurring and the recipient's aid in the search is being requested" (1986, p. 71). Thus a word-search gesture, which others have claimed aids the search, is here claimed to communicate a request for help.

One reason the communicative function of gesture is debated has to do with the definition of communication itself, as Krauss and Hadar (1999) pointed out. A common understanding of communication is that it is *intentional*. A number of researchers mentioned above have said, in effect, that "gesture is a visible part of the thought process". Because it is visible, gesture can depict thoughts, and thus can be claimed by some to be communicative. But if the speaker did not *intend* to communicate her gesturally-revealed thoughts, then others can claim the gesture was not communicative, but merely a cognitive byproduct (albeit a visible one). The question, then, is intentionality, which is difficult to pin down.

To conclude this summary of the debate, it should be noted that neither side maintains an all-or-nothing stance. McNeill (1992) specifically allowed for the occurrence of lexical-retrieval gestures (describing an example as grasping the air), even naming such gestures "Butterworths" in honor of their proponent. For their part, Butterworth and Hadar (1989) allowed that gestures do have a communicative function, but insisted that it is a secondary function. Similarly, Krauss et al. (1991) admitted some communicative function, such as Bavelas' conversation-regulating interactive gestures. In short, each side acknowledges the existence of the phenomena that the

other side holds up for evidence, yet denies its applicability towards a general model of the functions of gesture. It may well be the case that different types of gestures have different functions, as Krauss et al. (2000, p. 262) pointed out.

> Kendon (1996) summed it up this way:
>
> > There remains a controversy about the way in which gesture as an activity is related to speech. Some investigators appear to consider it simply as a kind of 'spill-over' effect from the effort of speaking, others see it as somehow helping the speaker to speak, yet others see it as determined by the linguistic choices a speaker makes as he constructs an utterance. An opposing view is that gesture is a separate and distinct mode of expression with its own properties which can be brought into a cooperative relationship with spoken utterance, the two modes of expression being used in a complementary way (see Kendon 1983). Careful studies of just how the phrases of gesture and the phrases of speech are related would throw useful light on this issue (cf. the recent dissertations of McClave 1991 and Nobe 1996).

One of the goals of my dissertation is to indeed "throw useful light on this issue" by carefully examining how the phrases of gesture and the phrases of intonation are related.

2.2 Intonation

Compared to the literature survey of gesture researchers, this review of intonation work will look relatively brief. This is not because intonation is a little-studied field—quite the contrary. But the focus of this dissertation is on the interaction between intonation and gesture. Gesture researchers, whatever their theoretical leanings, have sought to relate their field to speech. Yet the converse has not been true for linguists, who have not felt the need to concern themselves with bodily movement[7].

My review of the intonation literature, as it relates to gesture, is organized as follows. I'll survey a variety of researchers who have touched on both topics, however briefly. I'll then discuss the two researchers, both phonologists, who have studied intonation and gesture in depth: Bolinger and McClave. I will also briefly lay out the theories of another phonologist, Pierrehumbert, though she doesn't concern herself with gesture. I do this because the theoretical basis of the intonation side of my research is due to her, much as I rely on McNeill for the gestural side of my work. Finally, in the following chapter, I'll describe a pilot study I've carried out, investigating a subset of the intonation-gesture interface.

2.2.1 Researchers Touching on Both Modalities

In surveying the gesture literature above, I noted a number of gesture researchers who also considered intonation. I'll recap them here. Pike suspected a relationship between the two modalities. McQuown et al., Birdwhistell, Scheflen, and Condon looked at both modalities in support of interactional studies, and noted a synchrony of linguistic and kinesic stress. Kendon

[7] Exceptions, of course, are linguists who study sign languages. Interestingly, it might be thought that sign language linguists would not be concerned with intonation. Yet Sherman Wilcox (personal communication, June 11, 2001) and Scott Liddell (personal communication, March 13, 2002) have each mentioned to me the possibility that gradient modulations of sign forms may be a type of intonation. This fascinating kind of intonation, and its possible relation to both language and body movement, is unfortunately beyond the scope of my study.

aligned gestural and intonational hierarchies. Tuite noted a rhythmic pulse involving both channels. Finally, McNeill, Quek and colleagues correlated the modalities along the lines of discourse segments.

McNeill, Quek et al.'s study deserves further discussion in terms of intonation. As mentioned, they correlated gestural catchments to discourse segments. To identify discourse segments, they annotated prosody according to the ToBI (Tone & Break Indices) scheme (Beckman and Elam 1997; Silverman et al. 1992). Like other researchers (Grosz and Hirschberg 1992; Hirschberg and Nakatani 1996), McNeill, Quek et al. noted correlations between discourse segments and intonation. Deeply embedded discourse segments contained higher boundary tones, "conveying a 'more is to come' meaning", while more dominant segments contained lower boundary tones (McNeill et al. 2001b, p. 22). Because catchments were previously related to discourse segments, the authors concluded that "each catchment had its own distinctive boundary tone" (p. 22).

Relevant to my dissertation, McNeill, Quek, and colleagues did analyze gesture alongside an intonational framework such as Pierrehumbert's (upon which ToBI is based). However, they did not look at any elements other than boundary tones, nor units smaller than an intonational phrase, nor gestural units smaller than a catchment. What they really did was explore the alignment of three high-level entities: gestural catchments, discourse segments, and intonational phrases. The gesture/intonation relationship I'm investigating takes place at a lower level as well: inside both intonational phrases and gestural units.

Apart from the gesture researchers, recapped above, who have touched on intonation, there have been a number of others who have touched on both intonation and gesture, however briefly.

Starkey Duncan (1972) investigated the ways many devices, including gesture and intonation, are used for turn-taking. For instance, turn-yielding can be signaled by cessation of gesture, or by a rising or falling final contour. Duncan didn't compare gesture and intonation per se, but rather noted how each can be used as a conversational signal.

In a later study, Starkey Duncan and Fiske (1977) used intonation and gesture to help determine boundaries of "units of analysis" in support of face-to-face interactional studies. The intonational cue to such boundaries was the existence of a phonemic clause (containing one primary stress and one terminal juncture), per Trager and Smith (1957). Gestural cues to such boundaries were (1) hands at rest, (2) feet at rest, and (3) head turned away from the partner. Seven other cues were used, including syntax, lexical choice, and non-intonational prosody. Thus, the only relation which could be indirectly concluded between intonation and gesture was that terminal junctures tended to co-occur with hands at rest, feet at rest, and the head turned away. Duncan and Fiske did not use a modern framework from either intonation or gesture research. Intonation was annotated using Trager and Smith's (1957) system of four pitch levels and three terminal junctures. Gestural annotation noted general hand/body positions and movements, instead of recognized gestural units or types from any of the above-described gestural classifications. Finally, the basic unit on which all other events were annotated was not time, but syllables. Thus, every event was coded as happening either on or between successive syllables, and linear time was not accounted for.

Feyereisen and de Lannoy (1991, p. 94) discussed both gesture and intonation in terms of the emphasis each provides to the co-occurring speech. They said nothing further, however, about a relationship between the two modalities.

Benner (2001) compared the onset of gestures to the onsets of their lexical affiliate, and also to the tone unit in which the gesture occurred. She found these relationships sensitive to

narrative context. In narrative sections with more plot focus, intervals between gesture onsets and their subsequent lexical affiliates lengthened, presumably due to more complex gestures. Conversely, in these same narrative sections, intervals between gesture onsets and their preceding tone unit onsets were shorter, and the tone units tended to be longer. Some of the above effects were increased further if the gestures happened to be iconics or metaphorics, as opposed to deictics or beats. Relevant to this dissertation, the extent to which Benner analyzed intonation was the timing of tone units as a whole in relation to gesture. She, like Kendon, used the British School framework, and analyzed nothing smaller than a tone unit.

Creider (1978, 1986) also noted the alignment of gesture and stressed syllables. Making observations in several African languages, he described differences in the hand movements among the languages, differences which "appear to be conditioned by the nature of the use of stress in the prosodic systems of the languages" (1986, p. 156-7). In the language Luo, he found that beats were timed with the nuclear tone, and that a falling nuclear tone, in conjunction with a beat ending in a lowering hand or head, signaled the end of the speaker's turn. Given this correlation, he speculated that the turn-taking functions of tone groups and body movement were "at the very least closely related". He concluded, "If this is so then these body movements provide important evidence for our understanding of tone groups" (1978, p. 336).

Nobe (1996, 2000) refined Kendon and McNeill's phonological synchrony rule to a "gesture and acoustic-peak synchrony rule". He first defined an "acoustic peak" as being a peak of either F0 or intensity (or both, if they co-occur). His rule then stated that the gestural stroke precedes or co-occurs with (but does not follow) the *later* of the two types of acoustic peak (if they do not co-occur). Apart from locating the F0 peak in an utterance, and relating it to the gesture, Nobe did no further intonational analysis.

Valbonesi et al. (2002) compared the timing of gesture to speech on a larger scale, finding that gestures occurred during fluent phases of speech, and on a smaller scale, finding that strokes align with stressed syllables. Valbonesi et al.'s contribution was to determine focal points in speech automatically and reliably, using a variety of F0 and amplitude cues. Kettebekov et al. (2002) also used automatic detection of stressed syllables, which included using the pitch track, to improve automatic visual detection of gesture strokes. Their goal was to improve next-generation human-computer interfaces that can recognize gestures.

Ekman and Friesen, who as described above provided definitions for gesture types, also looked at the contributions various channels, including pitch and body movement, make during deception (Ekman et al. 1976). They found fewer illustrators during deception, higher pitch, and a negative correlation between the two (i.e. illustrators do not often co-exist with high pitch). This is the only relationship they studied between the two modalities. Furthermore, their method of analyzing pitch was incomplete, as "Pitch was measured by selecting two short speech samples from the subjects' answers and extracting fundamental frequency" (1976, p. 24). In other words, pitch was not measured throughout the data, but only in selected samples.

Scherer and Wallbott (1985), in a discussion of methodology for recording nonverbal behavior, showed an example of plotting F0 and hand movements on the same time axis. They said nothing, however, about analysis or the relationship between the two.

Finally, Cruttenden's 1997 textbook *Intonation* devoted several paragraphs (p. 177) to intonation and gesture, essentially recapping Bolinger's views (discussed in detail below).

To summarize the above work on intonation and gesture, many researchers have touched on certain features of intonation, but none have investigated gesture alongside a full framework of intonational phonology.

2.2.2 Bolinger

Bolinger (1982, 1983, 1986) agreed with the gesture community that gesture and speech (specifically, intonation) are a "single form in two guises, one visible and the other audible" (1986, p. 199). Indeed, he felt that "Intonation belongs more with gesture than with grammar" (1983, p. 157), and that "[features of pitch] *are* gestures … and that the contribution to discourse is of the same order" (1982, p. 19). His proposal for this commonality was that both channels are expressing the speaker's emotional state. He felt that intonation, like gesture, could be an "iconic", in this instance an iconic for the speaker's emotions. Like Scheflen and Birdwhistell, Bolinger suggested that pitch and body parts move in parallel—that is, they move up and down together. Bolinger's reason was that they both rise with emotional tension, and both lower with emotional relaxation. I will call this "Bolinger's Parallel Hypothesis".[8]

Bolinger gave a number of constructed examples in which the two modalities move in parallel. For instance (in his distinctive way of indicating pitch with typography), he wrote:

> Easiest to observe is the coupling of pitch with head movements. When a speaker says
>
> I
> wíll.
>
> … with a terminal rise, the head—if it moves at all—will move in parallel; to make it do the opposite requires practice. (1986, p. 200)

[8] Interestingly, McMahon (2003) claims that Optimality Theory (OT) is well-suited to one aspect of phonology—namely prosody—and not another—namely segmental phonology—because the two are quite different underlyingly. She feels prosody is evolutionarily older and more innate, and hence a good candidate for OT. To support her evolutionary claim, she cites (among other evidence) Bolinger's hypothesis linking intonation with (presumably innate) gesture.

Bolinger did provide a counter-example (1986, p. 200):

> One frequent exception to this covariation is the use of a downward thrust of the jaw to mark an accent … Since intonation marks accents much of the time with a rise-fall, pitch and head movement may go in opposite directions, e.g. in saying

I knó
w.

Bolinger also gave constructed examples where intonation and gesture combine in rich ways to signal a variety of meanings. For instance (1986, pp. 205-206):

> Take the following, which involves gesture at five levels besides intonation: head, eyes, eyebrows, mouth, and hands. The speaker is in an argumentative and rhetorical mood, and asks the … question

Does he néed ¡t?

> with *it* rising to falsetto. The hearer is expected to be compelled to say *no*, and that is supposed to clinch the argument. The rising intonation insists on a reply, and the concomitants are: (1) eye contact; this "holds" the listener to making the reply that the intonation insists on; (2) eyebrows raised; this is coupled to the intonation: high pitch, high eyebrows; (3) mouth left open, corners upturned; (4) hands out-flared, palms up: 'Everything is in plain view,' hence nothing is concealed, the case is obvious; (5) head shaking: 'The answer is *no*.' …. Each of these components contributes to the question and one can play with various combinations to test the contribution of each.

Bolinger went on to alter the components to make his point. For instance,

> … replace the smile with a "shrewd look"—the eyebrows are lowered and the nose is slightly wrinkled, removing the "openness" from the face; now the speaker is sharing a confidence: 'Knowing what you and I know, do you really think he needs it?'

Finally, Bolinger also noted that pitch accents tended to be high on "new" information to the discourse (Bolinger 1986, p. 79).

2.2.3 McClave

McClave's 1991 Georgetown Ph.D. dissertation is still the first and only full-scale empirical study of intonation and gesture. She analyzed several hours of filmed conversations, picking out 125 manual gestures for microanalysis. These she classified according to McNeill's typology. Along with the gestures, McClave also transcribed tone units, following the British School framework. Her criteria for tone unit boundaries (partially borrowed from Cruttenden 1986) were: pauses, pitch movements (typically falling at the end), phrase-initial anacrusis (rapid unstressed syllables), phrase-final lengthening, change in pitch direction (low then high) across boundaries, and changes in register at some new tone units.

McClave noted the boundaries of these tone units (to look for alignment with gestural phrases), their nuclei (to check for alignment with gestural strokes), and their terminal pitch direction (to test Bolinger's hypothesis that pitch parallels gesture direction). She used an interlinear tonetic transcription (the "tadpole" transcription), impressionistically transcribed, similar to Crystal and Quirk (1964).

McClave's findings both supported and refuted earlier proposals, as well as provided new insights. Contra Bolinger, she found no significant correlation supporting parallel directions of gesture and pitch. Supporting Kendon, she found that gestural phrases do align with tone unit boundaries. Of the 125 gestures she analyzed, only four crossed tone unit boundaries. (This also indirectly confirmed McNeill's observation that gestures rarely crossed *clause* boundaries). McClave also discovered that strokes and holds are shorter than normal when followed by others within the same tone unit. This 'fronting' of gestures suggests that speakers know in advance that they will express several gestural concepts along with a lexical concept, and time the gestures to finish at the same time as their verbal counterpart. This is further evidence that gesture is not a by-product of speech.

Supporting McNeill, McClave (1991, 1994) found that beats do coincide with tone unit nuclei (which are stressed). In contrast to all earlier researchers, however, she discovered that not all beats occur on stressed syllables. She found that beats seem to be generated rhythmically outward from the tone unit nucleus. That is, a "rhythm group" of beats exists anchored around the nucleus (on which a beat occurs), such that beats are found at even intervals from the nucleus, even if they fall on unstressed syllables or pauses. Even more surprisingly, this isochronic pattern of beats begins well *before* the nucleus, implying that the entire tone unit, along with its nucleus and the rhythmic pattern based thereon, is formed in advance, before the first word is uttered. McClave noted the similarities with Tuite's "rhythmic pulse". She also noticed, as did Condon, an "interspeaker rhythm", in which a listener produced beats during the speaker's utterance.

The above findings of McClave are all examples of speech influencing gesture. She found one phenomenon of gesture possibly influencing speech. In almost all cases, a gesture was delimited by tone unit boundaries. Yet in the rare cases where a gesture spanned several tone units, those tone units must have parallel intonational and stress patterns. It appears they would otherwise not be licensed to occur within the same gesture.

2.2.4 Pierrehumbert

As mentioned, I will briefly present the theories of Pierrehumbert, who provides the theoretical underpinnings for my intonational analysis. Pierrehumbert (Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Pierrehumbert and Beckman 1988) was among the first to provide a formal phonological account of intonation. Pierrehumbert described intonational contours as consisting of a linear string of high and low tones, combining to form the full intonational melody over an utterance. In addition to the phonological elements, phonetic interpolation operations take place to produce the surface contour.

There are three types of intonational events. *Pitch accents* are tonal movements attached to a stressed syllable. These can be a simple high tone, a simple low tone, or a bitonal combination of a high and low in either order. One tone will be aligned with the stressed syllable, and is starred in Pierrehumbert's notation. Thus, the six possible pitch accents (in English) are:

H* (a local peak at the stressed syllable)
L* (a local valley at the stressed syllable)
H+L* (a fall *to* the stressed syllable)
H*+L (a fall *from* the stressed syllable)
L+H* (a rise *to* the stressed syllable)
L*+H (a rise *from* the stressed syllable)

In addition to pitch accents, there are two types of *edge tones*, which may be either high or low. *Boundary tones* occur at the edges of intonational phrases. *Phrase tones* occur at the edges of intermediate phrases, which are smaller than intonational phrases. The existence of intermediate phrases in English has been the subject of some debate (see Ladd 1996 for a discussion).

3 Pilot Study

At the outset of my research, I performed a pilot study on one aspect of the relationship between gesture and intonation (Loehr 2001). Noting that researchers in both fields claimed a relationship to discourse structure, I was curious whether the two claims were related, and if so, how. (The pilot study also allowed me to explore methods for capturing, annotating, and analyzing digital video.)

The relationship of gesture to discourse structure I have already described. Among other things, McNeill claimed that beats signal a shift in narrative level. The shifts occur between "narrative" (telling a story), "metanarrative" (talking about the story), and "paranarrative" (stepping out of the storytelling). The utterance containing the narrative shift may contain several beats, sprinkled throughout the utterance.

The relationship between intonation and discourse structure has also been alluded to briefly. Bolinger noted that a high pitch accent will signal an item new to the discourse. This view was expanded upon by Pierrehumbert and Hirschberg (1990) (hereafter "PH"), and by Hobbs (1990).

PH argue, and Hobbs agrees, that intonation over some entity is used to describe the *relationship* between the entity and some other entity already in the discourse context. For example, an H* pitch accent indicates that the accented entity is *new*, in relation to what is already in the discourse context. The authors make no mention of the prediction going the opposite way; that is, they say nothing about whether a new item will require a high pitch accent.

Conversely, an L* pitch accent signifies that the accented entity is *not* new to the discourse. An analogy is the use of the indefinite (*a bus*) to introduce a new entity, versus the definite (*the bus*) to refer an existing entity.

The meaning of L+H* is slightly different. PH claim that L+H (in general, with no star) means that there exists some salient *scale* between the accented entity and some entity already in the discourse. L+H* (with the star) means, in addition, that the accented item should be put into the mutual belief of the discourse. This is most commonly used, they find, in corrections or contrasts. An example is (Pierrehumbert and Hirschberg 1990, p. 296):

A: It's awfully warm for January.
B: It's even warm for December
L+H* L H%

Here, A puts a chronological scale (months of the year) into the discourse. B then acknowledges that scale with an L+H, and then chooses an L+H* to state that the accented item (*December*) should be put into the discourse, updating the erroneous *January*.

Hobbs differs slightly from PH. He breaks down an L+H into its two components: L ("not new") and H ("new"). In his view, an L+H* is a "new" preceded by a "not new" qualifier. Thus, L+H* means "you might think this is *not* new (i.e. the month of the year, since you just mentioned it), but really it *is* new (since I just corrected your erroneous month)".

The contribution of intonation to discourse, then, is to signify the relationship (in our example, *new* or *not new*, and perhaps with a salient scale) of the accented item with the discourse context.

Having described the contribution of both gesture and intonation separately to discourse-level pragmatics, the question becomes: is there any correlation between these separate contributions?

To answer this question, I looked at the common site where each medium relates to discourse structure. On the gesture side, this meant looking at the *beats*. On the intonation side, therefore, this meant looking at the *tones* on the words where the beats occur. In my pilot study

data, did the beats and the tones associated with the beats convey discourse structure meaning? If so, what is the nature of such meaning?

At first glance, this may seem to be an apples-and-oranges comparison. The two modalities relate to discourse differently, both in terms of their effect and their scope. The effect of beats is to signal a narrative shift, while the effect of intonation is to signal the relationship of the accented item to an item in discourse. The scope of beats is across the utterance: a beat signals a narrative shift taking place somewhere in the *utterance* (not the *word*) in which the beat occurs. The scope of intonation's discourse contribution, however, is the accented item only (typically a word).

Such differences may not necessarily be a problem. It is conceivable that the two channels provide complementary information. In any event, the goal of the pilot study was simply to see if and how each medium conveys discourse structure information in real-world data, and then to see if there is any correlation.

My pilot study data was a 45-second segment of a popular televised talk show, "Regis and Kathie Lee" (Philbin et al. 2000). In the segment, a male speaker, Regis Philbin, is sitting between his female co-host, Kathie Lee Gifford, and their male guest, Matthew Broderick. Just prior to the segment, Matthew has been discussing his latest work in a Broadway play written by Elaine May, and Kathie Lee has expressed her admiration for the playwright. The transcript of the segment follows. For the reader's benefit, I have arbitrarily put utterance fragments on separate lines, included minimal punctuation, and described the more major gestures informally in square brackets.

Regis: You know, it's so funny Matthew that uh
I just happened to have dinner at a dinner party with Elaine May.
[deictic gesture towards Kathie Lee]

Kathie Lee: Of course you did.
[rolls eyes upward]

Regis: And so I said to her
<pause>
what's going on in your - first of all she didn't know who I was
[deictic to self]
<pause>
she had never she
<unrecognizable overlap with Matthew>

Matthew: She doesn't watch
<unrecognizable overlap with Regis>

Regis: Fifteen years in the morning
[deictic to floor]
doesn't
<pause>
doesn't know her
[deictic to Kathie Lee]
doesn't
<pause>
<unrecognizable>
to not to know her
[deictic to Kathie Lee]
that's really something
[hands to head in a "that's crazy" gesture]
but anyway
<pause>
<laughter>

Kathie Lee: <laughter>
Didn

Regis: Didn't know millionaire show didn't know anything!
[hands in side-to-side sweeping "negation" gesture]
<pause>

Kathie Lee: Didn't know millionaire show?
<laughter>

Regis: <laughter>
So-o so I said to her what're you doing I j-

all of a sudden she's telling me
[hands windmill in front of torso]
she's
<pause>
writing a movie
[deictic to Matthew] <pause>
she's writing a play
[deictic to Kathie Lee]
a Broadway play.
[repeat deictic to Kathie Lee]
So I said
<pause>
my gosh
<pause>
let me talk about this talk.
[changes posture to play role of Elaine May]
"Don't mention it
[defensive backing away motion]
<pause>
don't mention it!"
[defensive pushing away with hands motion]
<pause>
[changes posture back to play role of Regis]

Matthew: Oh really?

Regis: Two days later
[extends two fingers]
<pause>
out in the paper
[spreads hands to hold a newspaper]
Elaine's May
<pause>
Elaine May's
<pause>
play coming out. What is it, "Taller than a"
<pause>
"Taller than Dwarf?"

Matthew: "Taller than a Dwarf".

Regis displayed a rich variety of gestures. Deictics were abundant, towards all three participants. Iconics included gestures for a newspaper, and for the number *two*. A metaphoric was the windmilling motion to indicate that Elaine May was repeatedly telling him things.

Another one was pointing to the floor of the studio during *fifteen years in the morning*, to indicate he had hosted a morning television show for fifteen years. He used his whole body for postural shifts, even standing up slightly, to play the role of Elaine May. And his gestures accompanying *movie*, *play*, and *Broadway play* were all symmetrical, reflecting the syntactic symmetricality of their lexical counterparts. Finally, he even made beats on two stressed syllables during Kathie Lee's utterance *Didn't know millionaire show?*, confirming the interspeaker synchrony observed by Condon and McClave.

Although the gestures described above were interesting, my focus for this study was on the speaker's beats and their associated pitch accents. To explore this area, I used the following methodology. I should note that the tools for digital video capture, annotation, and analysis which I used for the pilot study differed somewhat from those I used for my dissertation research. I'll therefore only briefly describe my pilot study tools. The tools used in my dissertation research are fully discussed in the following chapter.

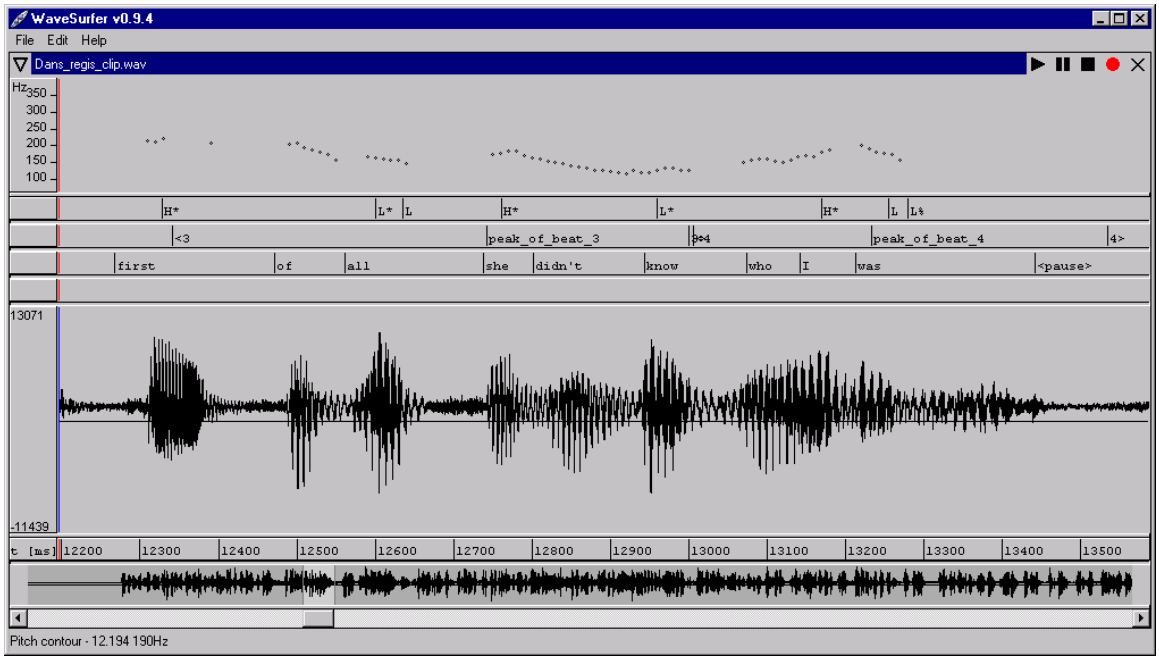
After recording the television video (at 30 frames/second) onto an analog tape using a video cassette recorder, I then digitized it with the Final Cut Pro software package. This software, designed for digital video production, allows one to step through a video frame by frame, or to mark and replay any segment repeatedly.

I first viewed the segment at full speed many times to absorb the gestures in real-time. I then stepped through frame-by-frame, searching for beats. I followed the methodology described in McNeill (1992, p. 380) for determining whether or not a movement was a beat. For instance, one criteria is that beats have only two phases, as opposed to the three phases of many other gestures.

Once the beats were identified, I noted their location (start, end, and point of greatest movement), in both video frame number and its equivalent time. I then analyzed the soundtrack

of the segment, using the acoustic analysis software WaveSurfer (Kjölander and Beskow 2000). WaveSurfer allows one to create parallel windows with waveform, pitch track, and transcriptions. Taking the timing marks of the beats from the video segment, I transcribed the beats onto a transcription window above the pitch track. I also transcribed the narrative itself. I was thus able to have time-aligned windows showing the waveform, pitch track, narrative, and beats. I then analyzed the pitch track surrounding each beat, and transcribed it according to Pierrehumbert's notation. In particular, I took care to mark the exact location of the pitch accent, using the waveform window to find the point of highest intensity in the syllable. Figure 1 shows a sample annotation screen, for the utterance "First of all, she didn't know who I was".

Figure 1. Sample annotation screen for pilot study.



At the top of Figure 1 is the pitch track. Below that is the intonational transcription, the beat transcription, and the lexical transcription. The vertical bars to the left of each label denote the exact point of reference. In the beat transcription track, two beats are shown, numbered 3 and

4. The start of each is labeled with a left angle bracket (e.g. "<3" denotes the start of beat 3). The end of each is labeled with a right angle bracket (e.g. "4>" is the end of beat 4). The end of beat 3 and the beginning of beat 4 almost exactly overlap (they are one video frame apart). The "peak", or point of greatest movement, of each beat is also marked. Notice how the peak of beat 3 precedes the H* over *she*.

For reference, I also kept the video software running on the computer screen, so I could easily see how a beat looked while reviewing the acoustic details. At the time of this pilot study, no software package existed that allowed both video and audio analysis in the same window. Since then, a satisfactory solution has appeared, which is to use the Anvil gesture analysis/annotation software (Kipp 2001) which imports audio annotations from the audio analysis/annotation package Praat (Boersma 2001). This newer solution is what I've used for my dissertation research, as described in the next chapter.

Once the beats and the intonational patterns on them were transcribed, I compiled the results into a table. I then checked for the validity of the claimed discourse contribution by each. Table 3 contains the results.

Table 3

Validity of gestural and intonational meaning in the pilot study

Gesture			**Intonation**		
Beat #	Accompanying word (italicized)	Is McNeill's claimed meaning valid? (Is there a shift in narrative level?)	Inton. on word	PH/Hobbs' claimed meaning of intonation on word	Is PH/Hobbs claimed meaning valid?
1	it's funny, Matthew, *that*, uh…	**yes**: from paranarrative to narrative (introducing story)	H*	"new"	**no**:*that* is not new item, not even an entity
2	Elaine *May*	**yes**: from narrative to paranarrative (using deictic to listener Kathie Lee to refer to earlier interaction with Kathie Lee)	L+H*	PH: "salient scale, put into mutual belief" Hobbs: "you might think it's not new, but really it is new"	**no**: *Elaine May* is already mentioned, but no salient scale present, nor anything new being said about entity
3	first of all *she* didn't know	**yes**: from narrative to metanarrative (backtracking)	H*	"new"	**no**: *she* is already mentioned
4	didn't know *who I was*	**yes**: ditto above	H*	"new"	**maybe:** *I* (speaker) not new item, but syntactic head *who* might be
5	to *not* to know her	**yes**: from narrative to paranarrative (leaving story to make comment about Kathie Lee)	H*	"new"	**no**: *not* is not a new item, note even an entity, and "not knowing" already mentioned
6	to not to know *her*	**yes**: ditto above	H*	"new"	**no**: *her* already mentioned
7	so-*o*	**yes**: from paranarrative to narrative (returning to story from shared laugh)	H*	"new"	**no:** *so* is not new item, not even an entity
8	let me talk about this *talk*	**yes**: shifting between characters in narrative	L*	"not new"	**yes:** this *talk* is not new to discourse
	Total yes:	**8 yes**			**1 yes**
	Total maybe:				**1 maybe**
	Total no:				**6 no**

There were eight beats found, all produced by Regis while speaking. (As mentioned, Regis also made beats while Kathie Lee was speaking. I considered such interpersonal synchrony out of the scope of this study.)

The second column of Table 3 lists the word accompanying the beat, in italics with immediate context. Note that beats 3 and 4 occurred in the same utterance, as did beats 5 and 6.

The third column notes the validity of McNeill's claim that beats signal shifts in narrative level. In all examples, this is indeed the case. In beat 1, the speaker (Regis) is beginning to tell his story. He's shifting from interacting with Matthew (paranarrative) to telling the first line of the narrative itself. In beat 2, Regis accompanies *Elaine May* with a nod and a deictic gesture to Kathie Lee, as if to say "who you were just talking about". This interaction with a listener is paranarrative. Beats 3 and 4 occur during a backtracking, with Regis leaving the story line to provide some earlier information (metanarrative). With beats 5 and 6, he leaves the story again, this time to interject a personal comment unrelated to the story. Beat 7 ends a shared laugh with his listeners (paranarrative), and shifts the level back to the storyline. Finally, beat 8 signals a shift between characters in the narrative (which McNeill also includes in his theory). In the narrative itself, Regis (at the dinner party) has just asked Elaine May if he may mention their talk. The very next utterance is intended to be that of Elaine May replying: *Don't mention it! Don't mention it!* The beat signals this embedded role shift.

As can be seen, McNeill's claims fare perfectly. It might be asked whether beats were so common that there always happened to be one on a narrative-level-shifting utterance. But there were no beats on utterances other than those shifting narrative levels. In the data, therefore, beats did indeed signal narrative level shifts.

How about the intonational claim of PH/Hobbs? Column 4 (the first column under the section labeled "Intonation") lists the pitch accent found over the word accompanying the beat. Six of the eight pitch accents are H*, which Pierrehumbert notes is the most common one. The other two are an L+H* (beat 2) and an L* (beat 8). The next column to the right lists, as a

convenience to the reader, the meaning of the pitch accent as claimed by PH/Hobbs. The final column notes the validity of this claim.

PH/Hobbs do not fare as well as McNeill. Six of the eight examples do not bear out their claim. In beat 1, the word *that* bears an H*. But it is clearly not a new entity to the discourse; it is not even an entity at all, but rather a complementizer. The word at beat 2, (Elaine) *May*, carries an L+H*. It is at least an entity, but I can think of no salient scale to satisfy PH. Nor does Hobbs' interpretation of "you might think it's not new, but really it is new" fit. Agreed, the entity *Elaine May* is not new, but Regis doesn't seem to be saying anything *new* about it. Similarly, the pronoun *she* in beat 3, is referring to *Elaine May*, old information. This would predict an L*, instead of the H* which occurs.

The next four beats also carry H*, predicting new entities. In beat 4, *who I was* could possibly signal a new entity, if one counts the wh-element *who* as a new entity that needs to be instantiated somehow syntactically. But in beat 5, *not* is not an entity. Even if the entity were the fact of "not knowing", this fact has just been mentioned, and is not new. Beat 6 is on *her*, (meaning Kathie Lee), old information mentioned in the previous sentence. Beat 7 is on another non-entity, the discourse marker *so*. Finally, beat 8 provides the only vindication for PH/Hobbs. The pitch accent is L*, predicting old information, and the accented item, the *talk* between Regis and Elaine May, has been the subject of the entire narrative.

Thus, the intonational theory of meaning, as analyzed so far, does not hold up in the data.

However, when we re-evaluate the intonational theory of meaning in the light of the *levels* of discourse that McNeill uses, it fares much better. Table 4 illustrates this.

Table 4

Validity of intonational meaning in the pilot study using levels of discourse

Beat #	Accompanying word (italicized)	Inton. on word	PH/Hobbs' claimed meaning of intonation on word	Is PH/Hobbs' claimed meaning valid using *levels* of context?
1	it's funny, Matthew, *that*, uh…	H*	"new"	**no**:*that* is not new item, not even an entity
2	Elaine *May*	L+H*	PH: "salient scale, put into mutual belief" Hobbs: "you might think it's not new, but really it is new"	PH: no Hobbs: **yes**: we were just speaking of her in one context (L), now I'm introducing her in the new paranarrative level (H)
3	first of all *she* didn't know	H*	"new"	**yes**: first mention of entity in new metanarrative level
4	didn't know *who I was*	H*	"new"	**maybe:** *I* (speaker) not new item, but syntactic head *who* might be
5	to *not* to know her	H*	"new"	**yes**: first mention of "not knowing" in new paranarrative level
6	to not to know *her*	H*	"new"	**yes**: first mention of entity in new paranarrative level
7	so-*o*	H*	"new"	**no:** *so* is not new item, not even an entity
8	let me talk about this *talk*	L*	"not new"	**yes:** this *talk* is not new to discourse
	Total yes:			**5 yes**
	Total maybe:			**1 maybe**
	Total no:			**2 no**

Table 4 retains the intonational columns of Table 3, yet changes the last column to note the validity of the intonational meaning using levels of discourse. In beat 1, the theory fares no better: the word *that* is still not an entity. And in beat 2, there is still no salient scale to satisfy PH. But Hobbs fares better here with his definition that "you might think this is not new, but really it is new". What Regis is indicating is that "we were just speaking of Elaine May (*not new, L*) in one discourse level (the narrative), but now I'm introducing her (*new, H**) in a new level (the paranarrative)". In Regis' paranarrative role of interacting with Kathie Lee, this is the first mention of Elaine May, and hence the H* part of the pitch accent is felicitous.

In three other beats as well, all with H*, does the intonational theory fare better using levels of discourse. In beat 3, Regis has just shifted to the metanarrative level, and *she* is the first mention of Elaine May in this new context. The use of an H* now is felicitous; it signals the introduction of the entity to the *new* level of the discourse structure. McNeill (1992) noted independently that often proper names are used, when pronouns would typically be expected, precisely in such situations when a new narrative level has been reached. Pronouns are used for existing entities, while proper names are used for new ones. Similarly, H* is used for entities that may not be new to the overall discourse, but are new to the discourse level just entered. Moving on to beat 5, Regis has just entered a paranarrative level, and this is the first mention of *not knowing*. Also in this new level, we have the first mention of *her* (this time meaning Kathie Lee) on beat 6.

Thus, whereas the intonational theory of discourse relation fared poorly on Regis' narrative at first sight (validated by only one out of eight examples), it fared better (five out of eight) when taking into account new discourse levels signaled by beats. It should be noted that PH/Hobbs' theory has nothing specific to say about the discourse context it uses. There is therefore nothing incompatible with using it with levels of discourse as McNeill does.

Thus, the contributions of gesture and intonation to discourse appear to be *complementary*, in the following way. Beat gestures signal that a different discourse level has been entered, and intonation signals how to interpret an entity in relation (*new* or *not new*) to the discourse level just entered.

The number of tokens used in this pilot study analysis is of course extremely small (only eight beats and eight pitch accents), so I can make no claims about the general validity of the above claim. But it is nonetheless an interesting trend.

More important than the findings of my pilot study, however, was the chance to successfully explore ways to digitally capture, annotate, and analyze gesture and intonation. This exploration bore fruit in the methodology used for my dissertation research, described in the next chapter.

This concludes my discussion of previous work in gesture and intonation. Before moving on to describing my dissertation research, it may be helpful to review the unresolved questions in each field, and restate my dissertation goals.

The gesture community, while in general accord (despite differing terminology) about the types of gesture, has not been in agreement about the function of gesture—whether it's for the benefit of the listener or the speaker. In the intonation community, the existence of Pierrehumbert's intermediate phrase in various languages has been debated. Perhaps a comparison of one modality with the other could shed independent light on each issue.

While a few studies have looked specifically at gesture and intonation, none has used Pierrehumbert's framework, nor have they used acoustic measurements. My goal, therefore, is to use Pierrehumbert's intonational framework, acoustically measured, to more closely examine the relationship between gesture and intonation. My hypothesis is that a careful look at the unit boundaries and unit types of each modality will reveal correlations in timing, structure, and meaning between the two channels.

More specifically, I'll ask five questions. First, do body movement and pitch move up and down together, as Bolinger hypothesized, reflecting increased or decreased emotional tension? Second, do the unit boundaries of each modality align? Third, do the unit types of each modality pattern together? Fourth, is there any correlation in meaning between the two channels? Finally, how do the respective rhythms of body and intonation relate?

To answer these questions, I'll now turn to my research, beginning with the methodology.

4 Methodology

In describing my methodology, I'll discuss the subjects, the filming, the annotation, and the post-annotation data processing.

4.1 Subjects

I recruited subjects who were native speakers of American English, and who did not know that the topic of study was gesture and intonation. They were told that the research had to do with "communication". The subjects were recruited in sets of pre-existing friends, so that in each filming, the group conversing was a pair or trio of good friends. This was done to help the conversation flow as naturally as possible. In addition, each group contained either all men or all women. The subjects were not paid for their time.

Fifteen subjects were filmed, in six groups. There were one pair and two trios of women, and two pairs and one trio of men. The number in each group was determined merely by the availability of suitable subjects who were friends.

Prior to filming, subjects were asked to sign a consent form (Appendix A), and were then asked to converse with their friends naturally, on any topic, for an hour. They were provided a list of suggested topics (Appendix B), but none of them chose to use it, conversing instead on subjects of mutual interest. This format of free-form conversation differs from many gesture experiments, in which the topic is constrained. My hope was to study natural conversation. In addition, this format was chosen by the only other empirical study of gesture and intonation, McClave (1991), and I wanted to replicate her conditions as much as possible.

After the filming, I selected short video clips for further analysis. I selected sections of conversation in which one person was primarily holding the floor for at least 30 seconds, while gesturing. I had initially hoped to select clips at random time intervals. However, the full annotation process was extremely time-consuming, taking well over an hour per second of data.

If a randomly selected interval were to contain no gesture (or even no speech), much time would be wasted. I settled, therefore, to selecting sections which clearly contained gesture. This, by the way, has been the method of nearly every gesture researcher to date, for similar reasons. The enormity of the annotation task also forced me to restrict my study to four clips, as described below, for a total of 164 seconds. This amount of data is in line with previous microanalytic studies. Kendon (1972) looked at 90 seconds, Condon and Ogston (1967) five seconds. McClave (1991) chose 125 gestures to annotate; my data contained 147.

Table 5 provides details of the data clips selected. The last clip was shorter than my initial criterion of 30 seconds, yet was chosen due to interesting interactional synchrony between the participants.

Table 5

Details of Data Clips Selected for Analysis

#	Name of data clip (conversational topic)	Date filmed	Duration	Speaker's sex and approximate age	Number of participants in conversation
1	"cupboards"	28 Sep 2001	61 seconds	Female, 50's	3
2	"musicians"	6 Jan 2002	44 seconds	Male, 40's	2
3	"drywall"	12 Jan 2002	37 seconds	Female, 30's	3
4	"sous-chef"	26 Jan 2002	22 seconds	Male, 30's	3

4.2 Filming

The filming took place in a private home. The first clip, somewhat of a pilot filming, took place in a kitchen; the others in a study. The speakers sat in portable director's chairs with armrests, and there was a separate video camera for each speaker. The arrangement of chairs and cameras for the first clip is shown in Figure 2. As can be seen, the cameras were in plain view.

Figure 2. Arrangement of chairs and cameras for first data clip.

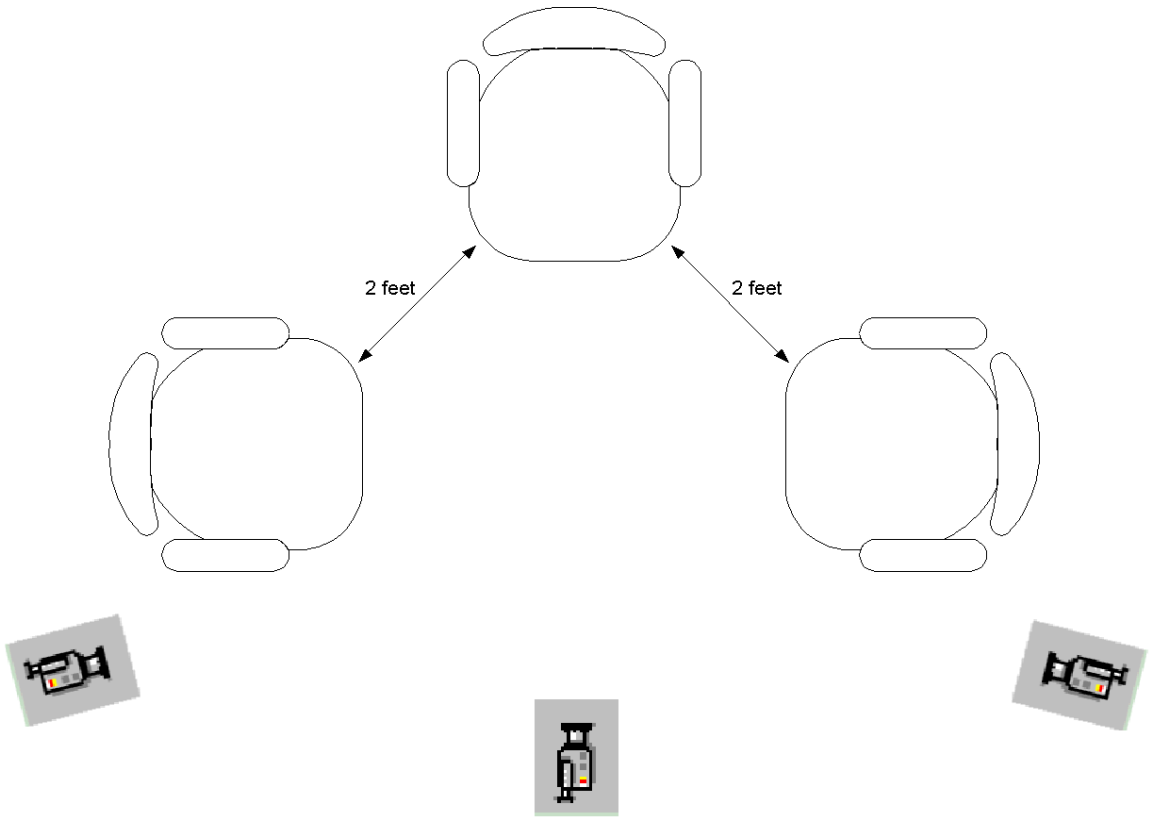
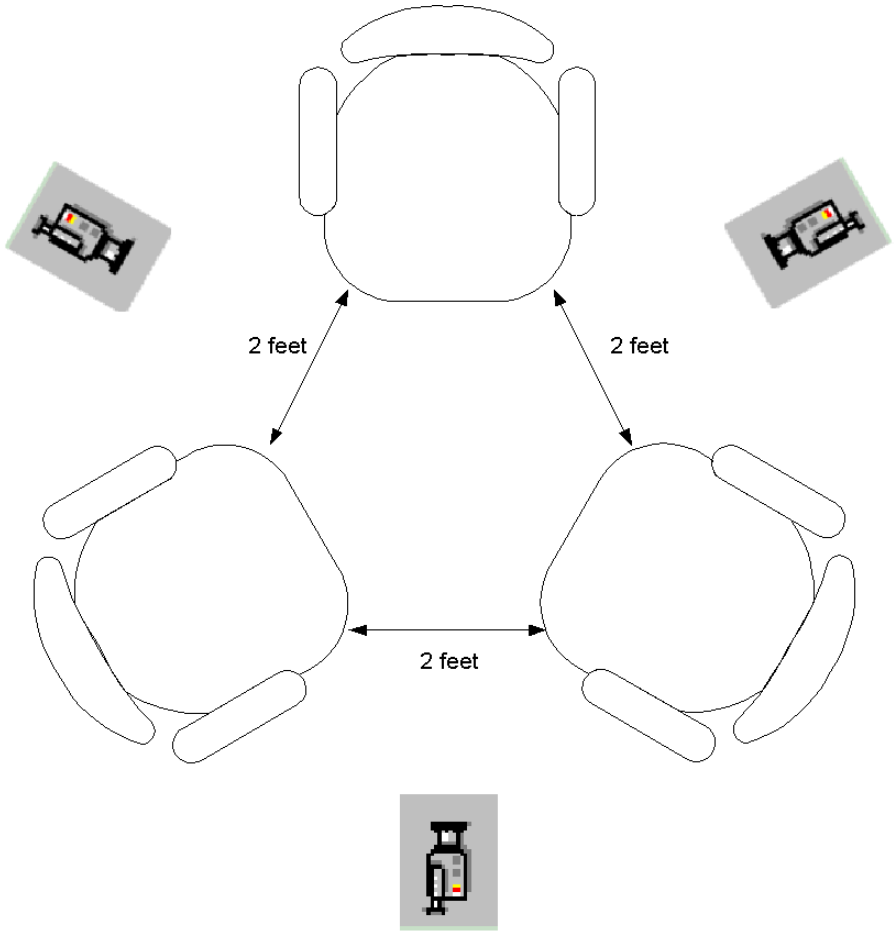


Figure 3 shows the arrangement of chairs and cameras for the remaining three data clips.

Figure 3. Arrangement of chairs and cameras for final three data clips.



One of the clips in the second arrangement had only two speakers. The three chairs were left in the same triangular arrangement, and one chair was left empty.

The video cameras used were Sony DCR-TRV30 Digital Video Camera Recorders, on tripods. Each camera was focused to capture a subject's body from the knees to the head. For the

first clip, the camera's built-in shotgun microphone was used to capture the speech of the speaker each camera faced. For the other clips, a lapel microphone was used (Radio Shack 33-3028 Omnidirectional Electret Stereo Microphone), which was plugged into the audio input jack of the video camera facing the microphone's bearer. Although the microphone collected stereo, only one channel was fed into the video camera. As the stereo microphone heads were less than an inch apart on the microphone, which was clipped to the speaker's lapel, no speech was lost by using a single channel.

Digital video was captured at 29.97 frames per second (fps), a standard North American rate, onto MiniDV tapes. One frame therefore equalled approximately 33 msec. The camera simultaneously captured audio, onto the same MiniDV tapes, at 48 KHz. The audio was more than suitable for subsequent pitch analysis.

The tapes were viewed using a digital video cassette recorder, and the four clips of interest were identified. These clips were then captured to a hard drive. An audio-only clip was saved to a .wav soundfile format, for intonational analysis. For gestural analysis, a video clip (with the audio track as well) was saved, in Quicktime format, with highest-quality Cinepak compression, at a window size of 360x240 pixels. The compression and smaller window size were necessary to digitally handle the otherwise huge data files, but the image quality was still good enough to see fine movements, including eye blinks.

As mentioned, the clips of interest each portrayed only a single speaker, with the other conversational participants out of view. Since the fourth clip had interesting interactional synchrony between the three participants, a digital video editing tool (Apple's Final Cut Pro) was used to create a multi-view video clip. This showed the videos of all three participants side-by-side, temporally synchronized, which allowed me to observe the movements of the three subjects together.

4.3 Annotation

I annotated the data in two stages. First, I annotated the audio for intonation, without reference to the video. Then, I annotated the video, without reference to the intonation annotations, but with the audio track playing during the video. It's possible to annotate intonation without video, but it's not possible to annotate gesture without audio in the McNeill style of annotation. I had originally hoped to annotate gesture without audio, to reduce bias, but soon found myself agreeing with McNeill and colleagues that the accompanying speech is crucial to interpreting and annotating gestures, as has been discussed at length in the literature review above.

Intonation was annotated in the Praat tool (Boersma 2001), and gesture in the Anvil tool (Kipp 2001). After both sets were completed, the intonation annotations were imported into Anvil, so they could co-exist with the gesture annotations, temporally synchronized.

4.3.1 Intonation Annotation

I coded intonation according to the ToBI (Tones and Break Indices) scheme, an implementation of Pierrehumbert's framework. I followed ToBI's published guidelines as described in Beckman and Elam (1997). ToBI has two levels of annotation: *Tones* (the *To* in ToBI), and *Break Indices* (the *BI* in ToBI). For tones, one marks the location and type of pitch accents, phrase accents, and boundary tones. Each tone is a high (H) or low (L), with additional symbols of '*' (for pitch accents), '-' (for phrase accents), and '%' (for boundary tones). Downstepped tones are marked with a preceding '!'.

For break indices, one denotes levels of juncture between words. There are five break indices, as follows, in increasing amounts of juncture:

Level 0: Minimal juncture, e.g. inter-word co-articulation such as flapping

Level 1: Used for typical inter-word boundaries

Level 2: A compromise used when pausing indicates a phrase boundary but intonation doesn't, or vice versa

Level 3: Used for a juncture corresponding to a intermediate phrase boundary

Level 4: Used for a juncture corresponding to an intonational phrase boundary

As I was only interested in the tones delimiting intermediate phrases and intonational phrases, I only marked the highest two break indices, levels 3 and 4, corresponding to intermediate phrase boundaries (phrase accents), and intonational phrase boundaries (boundary tones), respectively.

In Praat, I first used the spectogram to help identify word boundaries, transcribing the speech along the way. Then I used Praat's autocorrelation algorithm to generate a pitch track. The algorithm isn't foolproof, even when the window size is adjusted to individual speakers, and there were several sections where pitch halving or doubling occurred. I obtained the accurate pitch for these sections by manually counting glottal pulses per unit time.

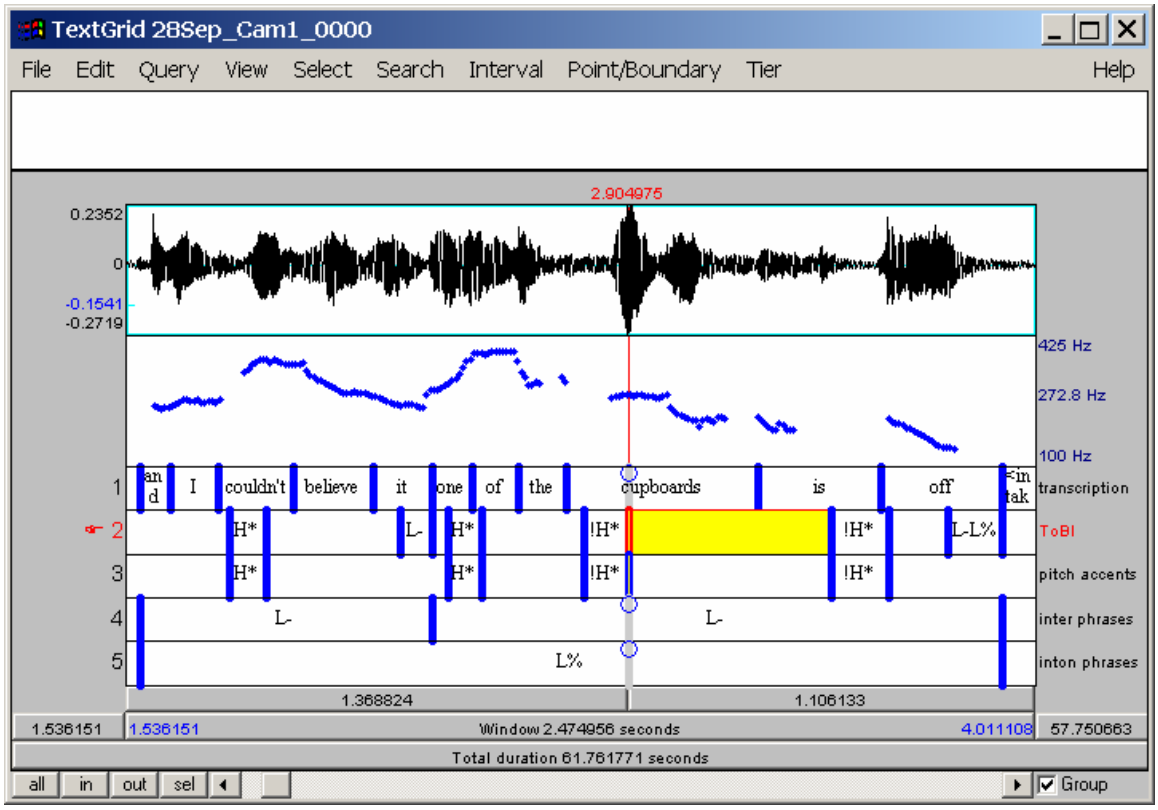
Once the pitch track was ready, I identified the locations and types of tones. Locations for edge tones (phrase accents and boundary tones) are defined by ToBI to coincide with word boundaries, so these tones were placed at the end of the phrase-final word[9]. Pitch accents are placed at the point of highest (or lowest) pitch within the vowel of the associated stressed syllable. I used the waveform amplitude to help determine this point.

[9] My data contained no phrase-initial edge tones.

I determined the types of tones based on the ToBI annotation guidelines, and also on the numerous examples (with audio files) in the associated ToBI training manual. Figure 4 displays a sample of ToBI annotation in Praat, for part of the "cupboards" data clip.

In Figure 4, the top two tiers display the waveform and pitch track, respectively. The third tier displays the transcription and word boundaries (the vertical bars surrounding the words). The fourth tier contains the standard ToBI marking of tones and their locations. Each tone is placed at a single point in the timeline, denoted by the vertical bar to the *right* of the annotation. For instance, in the first syllable of *cupboards*, there is a downstepped high pitch accent, labeled *!H**. The exact point of this tone is at the vertical bar to the right of *!H**. The bar to the left is simply a meaningless convenience, to keep the label near the right-hand bar for readability (Praat doesn't allow right-justification of labels). Also notice on this tier that phrase accents and boundary tones are often written together, if they occur together. The rightmost label, coinciding with the end of the word *off*, is *L-L%*, which indicates that a low phrase accent and low boundary tone finished the phrase.

Figure 4. Sample view of ToBI annotation in Praat, for the "cupboards" data clip.



To identify the tones, I relied heavily on ToBI's published training manual, complete with numerous soundfile examples. The most common tone, and the easiest to discern, was a simple H*. This is a simple local high, with no steep rises or falls around it. The single-toned L* was much rarer, and its low target stood out less clearly from the surrounding pitch track than the high target of an H*, perhaps because speakers have less frequency-room to maneuver in at the bottom of their range. This is an appropriate place to mention that ToBI allows for underspecification of tones. That is, if an annotator is sure that, say, a pitch accent is present, but is unsure what *type* of pitch accent it is, the annotator may use an *X** to denote this. In this way, at least the existence and location of the pitch accent are captured, even though the type is not. Underspecified phrase tones and boundary tones are labelled *X-* and *X%*, respectively.

Bitonal pitch accents were also less common. The distinguishing feature was a steep rise to, or fall from, the stressed syllable, a rise which could both be heard impressionistically and seen visually in the pitch track. There was a grey area, of course, between the shallow curves around a simple pitch accent and the steeper curves around a bitonal one, and not all decisions were easy. I'll get back to this point shortly when I discuss inter-annotator reliability.

As for phrase accents and boundary tones, these were often quite clear, both in terms of location (based on juncture) and type (based on the pitch contour at the juncture). Again, I followed ToBI's examples carefully. Following Pierrehumbert, simple (and common) declarative phrase-final falls received the L-L% combination of a low phrase accent and low boundary tone. The rarer question-rise received the opposite H-H%. Somewhat common plateaus received an H-L%; by Pierrehumbert's definition, the high phrase accent (H-) causes "upstep" on the low boundary tone (L%), resulting in a plateau.

As for phrase accents by themselves, the type was usually easy to discern. Simple pauses typically had a slight fall; these received an L-. Short pauses with a "continuation" contour, in which the speaker clearly had more to say, received an H-. The difficulty with phrase accents was not so much in deciding the type, but rather in discerning their existence in rapid speech—that is, was there an intermediate phrase boundary or not? As I'll discuss, I was grateful to have my annotations checked by another annotator, to be confident that the annotations I used in my analysis were reliable.

Returning to Figure 4, recall that the ToBI tier, just below the transcription, contains all ToBI tone markings together. The next three tiers denote pitch accents, intermediate phrases, and intonational phrases individually. The markings on the pitch accent tier are no different than their counterparts on the ToBI tier. Yet the tiers for intermediate and intonational phrases denote the *intervals* that these phrases cover, not the edge tones at the end. A phrase accent denotes the *end*

of an intermediate phrase; the intermediate phrase tier denotes the *span* of the intermediate phrase, from its beginning at the start of the first word of the phrase, to its ending at the end of the final word in the phrase. The intonational phrase tier is similarly arranged. In these two tiers, then, the vertical bar to the left of the label is meaningful; it denotes the start of the phrase, while the right-hand bar denotes the end of the phrase. Note in Figure 4 that there are two intermediate phrases (ending in *L-*) within one intonational phrase (ending in *L%*).

My annotations were checked against those of an experienced ToBI annotator[10], who independently annotated all the sound files in their entirety. I then compared the two annotations, and calculated inter-annotator agreement percentages (Table 6), following the procedure put forth by Pitrelli et al. (1994), who carried out a large-scale evaluation of ToBI inter-annotator reliability.

[10] I thank Katy Carlson for her generous help.

Table 6

Inter-annotator agreement for ToBI annotations

	Pitch Accents		**Phrase Accents**		**Boundary Tones**		**All Tones**		**Notes**
	#	*%*	#	*%*	#	*%*	#	*%*	
Number of words in data	525		525		525		1575		
Agreement on existence of tone type	494	***94%***	489	***93%***	484	***92%***	1467	***93%***	*Number of words in which both transcribers agreed on the existence of a certain tone type or not (e.g. both agreed that there was a pitch accent on that word, or both agreed that there was a phrase accent, etc.) Divided by above row to get percentage.*
Then, agreement on exact tone	462	***94%***	477	***98%***	483	***100%***	1423	***97%***	*Of those words where both transcribers agreed there was a certain tone type, this is the number where both transcribers agreed on the exact tone (e.g. they both agreed that there was an H*, or a L-, etc.) Divided by above row to get percentage.*
Absolute agreement		***88%***		***91%***		***92%***		***90%***	*Number in above row, divided by total number of words, to get percentage agreement on both tone type existence and exact tone.*

Note. All percentages rounded to the nearest whole number.

In Table 6, percentages are based on the number of words in the data. First, for each word, did the two annotators agree whether or not a certain tone did or did not exist on that word? For example, in the "Pitch Accent" column, did they agree that there was a pitch accent on that word? In the "All Tones" column, did they agree that there was a tone of any kind on that word, regardless of type? The number of agreements, divided by the total number of words, produces the percentage of agreement on tone existence. The second data row in Table 6 shows the number and percentage of these agreements, for each type of tone, and for all tones combined. There was 93% agreement on existence of tones, for all tones combined.

The next percentage is based on those words for which the transcribers agreed a certain type of tone existed. Of these words, on how many did the transcribers agree on the exact tone[11]? For example, given that they've agreed that there is a pitch accent, do they now agree that there is a H*? The third data row in Table 6 shows the number and percentage of these agreements. There was 97% agreement on the exact tone, given previous agreement on tone existence.

The final percentage is absolute agreement. This is the number of words for which the transcribers agreed on the exact tone (and hence on existence), divided by the total number of words. The last row in Table 6 reports these. The bottom line is that there was 90% absolute agreement, for all tones combined.

These inter-annotator figures compare very favorably those of Pitrelli et al. Table 7 compares absolute agreement between the two studies.

Table 7

Inter-annotator absolute agreement for Pitrelli et al. (1994) and for present study.

	Pitch Accents	**Phrase Accents**	**Boundary Tones**	**All Tones**	**Notes**
	%	***%***	***%***	***%***	
Pitrelli et al. (1994)	***68***	***85***	***91***	***81***	26 transcribers
Present study	***88***	***91***	***92***	***90***	2 transcribers

The probable reason why the present study has such high percentages compared to Pitrelli et al. is that the latter figures were calculated on pair-wise agreements between 26 transcribers, with much more chance for disagreement than between the mere two transcribers in

[11] Following Pitrelli et al., L+H* was counted as matching H*, as these two pitch accents are difficult for transcribers to reliable distinguish. Similarly, L+!H* and !H* were allowed to match each other. But tones were not allowed to match their downstepped counterparts; e.g., H* was not allowed to match !H*.

the present study. Nevertheless, the ToBI annotations in the present study are shown to be quite reliable.

In spite of the reliability of my transcriptions, however, there is still a concern. As my analysis (and eventual findings) depend on correlations of unit boundaries, it's crucial that my unit boundary annotations be of high quality. Pierrehumbert emphasized the importance of this, cautioning that any analysis of intonation alongside gesture should be done on phenomena with solid inter-transcriber reliability (personal communication, June 18, 2002).

Yet as I've explained, the placement of phrase accents (delimiting intermediate phrase boundaries) was the most difficult of my decisions. Pierrehumbert herself has acknowledged that some ToBI distinctions are difficult (ibid). Even though the inter-annotator agreement figures for the existence of phrase accents was good (93%, as shown in Table 6 above), that number is misleadingly high. It's misleading because the denominator in these calculations is the *total number of words spoken*, not the number of words with a likely phrase accent on them. As most words do *not* have a phrase accent on them, it's easy for two annotators to agree whether or not a phrase accent exists on any given word. Most words don't have a phrase accent, and both annotators agree on that, and right off the bat their inter-annotator agreement figure is high. The real question is: of those words on which at least one annotator thinks a phrase accent exists, then how well do they agree?

Using this more stringent criteria, my inter-annotator agreement figures fall to around three-fourths, for existence. For example, when looking at phrase accents, instead of starting with the 525 words in my data, I started with the 143 words on which at least one annotator felt a phrase accent existed. Of these, both annotators agreed that a phrase accent existed on 111 words, for a 77% agreement on existence of phrase accents. Since, as will be seen, my findings

depend crucially on the placement of tones, I therefore chose to be conservative and include for analysis only those tones which both annotators felt existed.

The above decision was based on tone *existence*. Although, as will be seen, I had no significant findings based on the *exact* tone (e.g. H*, L-), I likewise included for such analyses only those tones for which both annotators agreed on the exact tone.

I've taken care to explain my inter-annotator reliability figures, and my decisions based on them, so the reader can judge the soundness of the data on which I've based my findings. My intent was to make my data as reliable as possible.

4.3.2 Gesture Annotation

When the intonation annotations were done, I set them aside, and turned to gesture annotation, using the Anvil tool. I annotated according to the guidelines published by McNeill (1992), and amplified by Susan Duncan in both written instructions (2002) and in a two-day tutorial at the University of Chicago. I deviated from the McNeill guidelines in several ways, as I'll explain below.

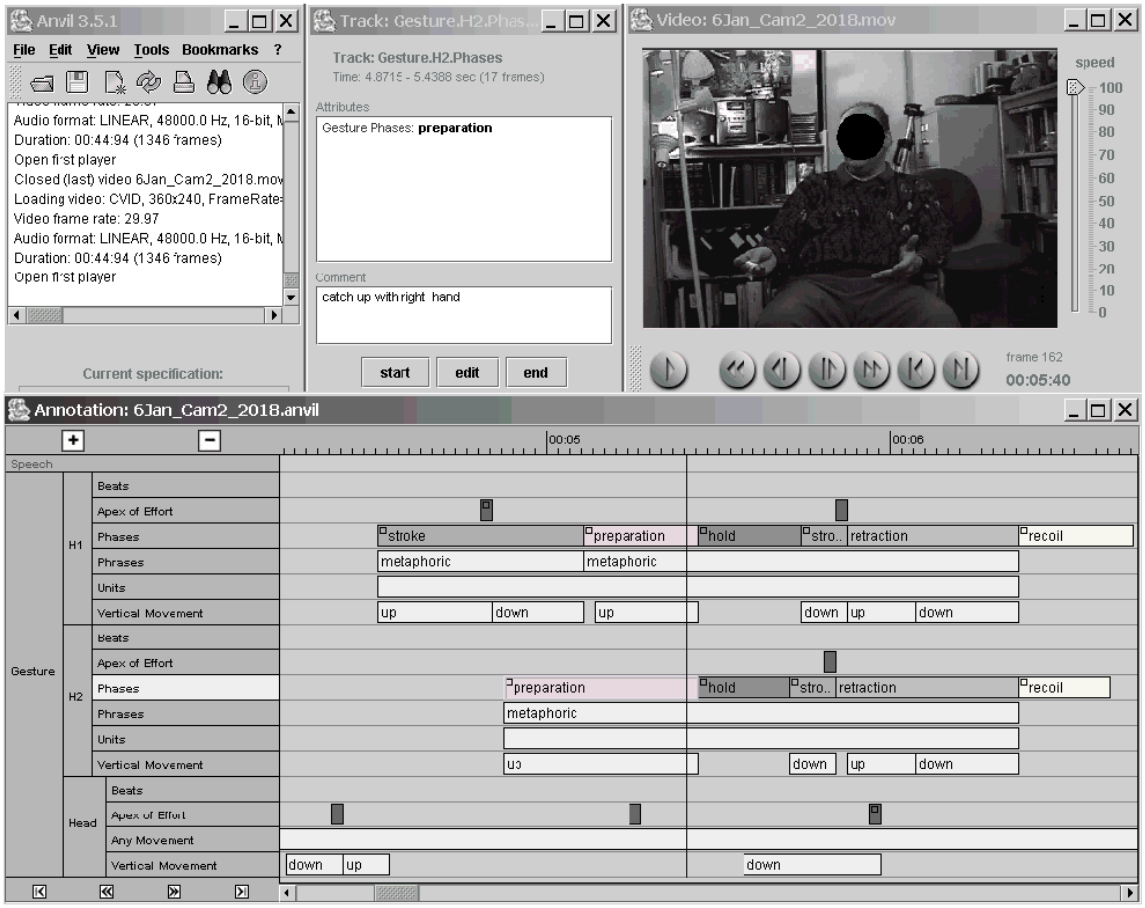
Figure 5 shows a sample annotation view in Anvil, of a portion of the "musician" data clip. The following three paragraphs describing Anvil's user interface are paraphrased from Loehr and Harper (2003).

The upper right window displays the video, and contains playback controls including variable playback speed and single-frame movement. (The subjects' faces are masked in this and subsequent figures to protect identity). The upper left window provides details of program execution. The upper middle window gives information about the currently selected track and annotation element (described below).

The main window at the bottom is called the *annotation board*, and is laid out like a musical score. The horizontal dimension is time, in units of video frames (note 30 tick-marks between each full second). The vertical dimension is a collection of *tracks*, each containing its own user-defined annotation type.

There is also a vertical playback line, running across all tracks. The playback line is synchronized with the current video frame. As the playback line is moved forward or backward, the video follows suit, and vice versa. This facilitates annotation. The user can click in a track at the starting frame of an interval of interest (e.g. a stroke), advance the video as slowly as desired to the ending frame of the interval, and use menu shortcuts to mark the appropriate item type with the given endpoints.

Figure 5. Sample view of gesture annotation in Anvil, for the "musicians" data clip.



I'll describe the specific gesture phenomena I annotated by going over the annotation tracks shown in Figure 5.

The topmost annotation track is labeled *Speech*. This track was kept empty during the gesture annotation, but afterwards contained the annotations from Praat, which were subsequently imported into Anvil. More precisely, *Speech* is a grouping of tracks (Anvil allows tracks to be grouped for convenience). The group *Speech* is "collapsed" in Figure 5, meaning that its component tracks (speech transcription and three levels of intonation) are temporarily hidden. Temporarily collapsing certain groups allows the user to better visualize other groups of interest.

The other main grouping is labeled *Gesture*, and it contains the gesture annotations, which I'll now describe.

The first six tracks in the *Gesture* group belong to a subgroup called *H1*, which stands for *Hand 1*. Hand 1 refers to the speaker's dominant gesturing hand, if there was one. If both hands were involved in a two-handed gesture, then that gesture was recorded in the Hand 1 group as well.

The first of the Hand 1 tracks marks *beats*. I used McNeill's "beat filter" criteria (1992, p. 380) for determining beats; the primary criterion is that beats have exactly two phases (typically an up-down or a down-up movement), as opposed to the usually at least three phases of other gestures. Since beats can be superimposed upon other gestures, I reserved a separate track for beats above the tracks of the other gestures. Annotations for beats consist simply of their starting frame and ending frame. There are no beats shown in Figure 5.

The next track marks what I call the *apex of movement*. Kendon defined the stroke of the gesture as the "peak" of the gesture. The stroke is an interval of time. I was interested not only in the the stroke—an interval—but also in the single instant which could be called the "apex" of the stroke, the "peak of the peak", the kinetic "goal" of the stroke. I wanted the apex for more precise timing information vis-à-vis speech, since strokes often last a number of frames. As I'll discuss later, this proved to be fruitful, as apexes turned out to align closely with the apexes of other articulators, and with pitch accents.

Coding the apex was the first of my deviations from McNeill's guidelines. If the stroke was uni-directional (e.g. a deictic), the endpoint appeared to be the "goal" of the stroke, and I marked the stroke's final frame as the apex. If the stroke was bi-directional, the point where the directions changed appeared to be the most important point dynamically, and this I marked as the apex. Multi-directional strokes received multiple apexes. I coded apexes on beats as well, by the

same criteria. I used a single track for apexes of both beats and strokes, since beats and strokes never had simultaneous apexes.

The apex occurs at an instant in time, yet the apex track in Figure 5 shows apexes with a durational interval of exactly one frame. This frame represents the closest frame to the instant in time when the apex occurred. Like pitch accents (which are also points in time), the right-hand boundary of an apex annotation represents the edge of the frame closest to the actual point of the apex.

The track below the apexes contains gestural phases, or g-phases, as defined by Kendon. These include preparations, strokes, retractions, and holds. These were decided upon based on criteria in McNeill (1992, pp. 375-376). Holds were simple to identify; the hands were motionless (or nearly motionless) while in the air. Strokes were determined both kinetically (the phase with the most "effort") and semantically (the phase containing the "content"). Preparations and retractions were, like holds, easy to identify, as the hands are either leaving or returning to rest.

The boundaries between strokes and preparations or retractions were typically readily discernible, thanks to either intervening holds, or marked changes in hand direction or form. There were occasional fuzzy boundaries, but these were not the norm. For any g-phase following motionlessness (e.g. following a rest position or a hold), I marked the beginning of the g-phase at the first sign of movement. Similarly, I marked the end of motion g-phases at the last sign of movement before stillness.

All in all, I found annotation of gesture units more straightforward than annotation of intonation units, perhaps because it's easier to visually discern changes in movement.

I also marked recoils, another change from McNeill. Recoils typically occurred after a hand finished retracting to a rest position, and consisted of a slight movement in which the hand

"bounced" back up slightly before finally settling down. I considered recoils "outside" of a gesture, yet since one goal was to faithfully record every movement of the hands, I marked them.

The next track contains gestural phrases, or g-phrases, which are collections of the above g-phases. Kendon defined a g-phrase as containing an obligatory stroke, with an optional preparation, and/or retraction, and/or hold(s) before or after the stroke. The endpoints of the g-phrase coincide with the endpoints of the outermost g-phases contained by the g-phrase. If two successive strokes were separated by a hold, the hold was judged to be a post-stroke hold, per McNeill, rather than a pre-stroke hold. Therefore, the g-phase containing the first stroke also contained the hold.

G-phases have types (e.g. deictic, metaphoric), which I noted. Following McNeill, I also noted the *meaning* of the g-phase, as best as I could determine. Due to space limitations, Anvil doesn't display all user-defined attributes (such as gesture meaning) on the annotation board, but they can be retrieved with a menu, and are exported along with all other annotation information.

Deviating from McNeill, I did not code attributes such as hand shape and hand position for each gesture. It's appropriate to record such attributes when the annotations exist *separately* from the original video. But, as pointed out in Loehr and Harper (2003), digital annotation tools such as Anvil allow annotations to *co-exist* with the original video footage, and such attributes can be readily seen in the video window. Therefore, it wasn't necessary for me to code hand shape and position[12].

[12] A reason to explicitly annotate hand shape and position would be if one were to analyze those features statistically, by exporting the annotations. I was interested rather in the timing of gesture boundaries, and in the types of gestures (e.g. iconic), for comparison with intonation.

The track below g-phrases contains gestural units, or g-units. These are defined by Kendon to contain all g-phrases, from the point where the hands leave a resting position, to the point where they return to one. In Figure 5, the g-unit shown for Hand 1 contains two (metaphoric) g-phrases. G-units do not have types; they are simply intervals, whose endpoints match the endpoints of the outermost g-phrases contained by the g-unit.

The final track in the Hand 1 group contains elements of vertical movement, specifically recorded to test Bolinger's parallel hypothesis. I marked the endpoints and direction (up or down) for any intervals with predominant vertical direction.

The next six tracks belong to the group H2, or Hand 2. These tracks, which are identical to those of Hand 1, were used if the non-dominant hand was performing a movement not part of a two-handed gesture. The Hand 2 tracks were not often used, as the non-dominant hand usually either participated with the dominant hand in a two-handed gesture, or remained at rest. (I did often notice Hand 2 performing adaptors, which I marked. My criterion for marking adaptors was a movement in which the hand manipulated an object or part of the body). In Figure 5, although the latter part of the Hand 2 gesture is clearly parallelling Hand 1, the beginning of Hand 2's gesture is quite different, starting much later, and not mirroring Hand 1's initial stroke. For this reason, I coded a separate gesture for Hand 2.

The final gesture grouping is for the head. Three of these tracks—beats, apex of movement, and vertical movement—are identical to those used for the hands. I chose not, however, to annotate Kendon's hierarchy of g-phase, g-phrase, and g-unit for the head. This is partly because I initially thought to constrain my research to the hands, but soon discovered that to explore theories of both rhythm and of Bolinger's parallel hypothesis, the head was indispensable to the equation. I therefore coded those aspects having to do with rhythm (beats and apexes) and with vertical movement. The other reason I ignored Kendon's hierarchy here

was that, although the head certainly performs gestures as surely as the hands do, it was less clear to me (and was not discussed in McNeill's guidelines) what constitutes a preparation, retraction, hold, or resting position for the head. I therefore simply marked intervals of *any* movement of the head, as shown in the second track from the bottom. In Figure 5, the head is in slight but constant motion throughout the annotation segment in view.

Since I didn't code Kendon's hierarchy for the head, I didn't include head movements in my analysis for three of questions: alignment of gestural and intonational unit boundaries, correlation of gestural and intonational unit types, and relationship between gestural and intonational meaning. For those questions, I only looked at hand gestures. Head movements were included in my analysis of rhythm, and in investigating Bolinger's parallel hypothesis.

The annotations for one of the data clips (the "drywall" clip), or approximately 25% of my annotations, were checked against those of an experienced gesture annotator[13], who independently annotated the data clip. I then compared the two sets of annotations.

The gesture community has never done a formal inter-annotator reliability study, with published procedures, metrics, and agreement figures (although one is currently being initiated (Susan Duncan, personal communication, February 25, 2004)). Without a community consensus on metrics, I chose to use a rather stringent one similar to the one I used for intonation annotation. That is, of all the places in which *either* annotator felt a gesture event took place, how well did they agree? For instance, if annotator A felt a g-phase or g-phrase began on a certain word, did annotator B mark likewise? Did they agree on the ending word of the same g-phase or g-phrase? For those they agreed on, did they then agree on the type (e.g. stroke or hold for g-phase, iconic or metaphoric for g-phrase)? For g-phrases, did they agree on the meaning of the gesture?

[13] I'm indebted to Susan Duncan for her generous help.

Table 8 shows the inter-annotator agreement for both boundaries and types, of both g-phases and g-phrases.

Table 8

Inter-annotator agreement for gesture annotations.

	G-phases		**G-phrases**			**Notes**
	G-phase boundaries	G-phase types	G-phrase boundaries	G-phrase types	G-phrase meanings	
Number of annotations marked	136	52	52	21	18	*Combined set of annotations marked by either annotator*
Number of annotations agreed upon	103	46	42	21	16	*Of above, number of annotations which both annotators agreed upon*
Percentage of annotations agreed upon	76 %	88 %	81 %	100 %	89 %	*Above row divided by top row*

As can be seen in Table 8, there were 136 g-phase boundaries marked in total (68 g-phase starts, and 68 g-phase ends). Of these, both annotators agreed on 103 boundaries, for an agreement percentage of 76 %. The criterion for boundary agreement was whether both annotators marked the boundary on the same word or on an adjacent word (as some g-phase boundaries occurred near word boundaries). Occasionally, agreement was allowed if the two annotations were separated by an intervening word, but in all such cases the intervening word was a rapidly spoken function word, and the distance between the annotations was less than 200 msec. As will be shown in the next chapter, 200 msec is much less than the typical word length in my data.

The boundary discrepancies between the two annotators were typically due to one annotator marking two shorter g-phases which fit within one longer g-phase marked by the other annotator. In other words, one annotator inserted multiple or "extra" phases which the other annotator didn't. For this reason, there were only 52 g-phases whose existence both annotators

agreed upon, as shown in the top of the third column in Table 8. Of these, the annotators agreed 46 times on g-phase type (e.g. stroke, hold), for 88% agreement.

The same strategy was used for calculating agreement for g-phrases, as shown in the next two columns. There was 81% agreement on boundaries, and 100% agreement on type (e.g. iconic, metaphoric). In addition, there was 89% agreement on gesture meaning[14] (e.g. "walls coming down", or "conduit metaphor"). As was the case for g-phases, the discrepancies in g-phrase boundaries were due to multiple phrases marked by one annotator, in place of a single phrase marked by the other. There were thus no phrases marked in a location where the other annotator had no annotations at all.

The agreement percentages, while not perfect, are still fairly high, given the stringent metric used. As an aside, the lowest-scoring category, g-phase boundaries, turned out not to be used in any of my findings, as will be discussed in the next chapter.

This concludes the description of exactly what I annotated, and how.

Later, when specifically looking at rhythm, I also decided to annotate eyeblinks, merely marking the interval from when the eyelids started to close to when they returned to being fully open. The subject in the "musicians" clip also made rhythmic leg movements for several seconds, which I chose to record. I'll show examples of both of those annotations when I discuss rhythm in Chapter 5.

[14] Three of the 21 g-phrases were left unspecified for meaning by one of the annotators. The 18 remaining g-phrases were used in calculating the agreement percentage for meaning

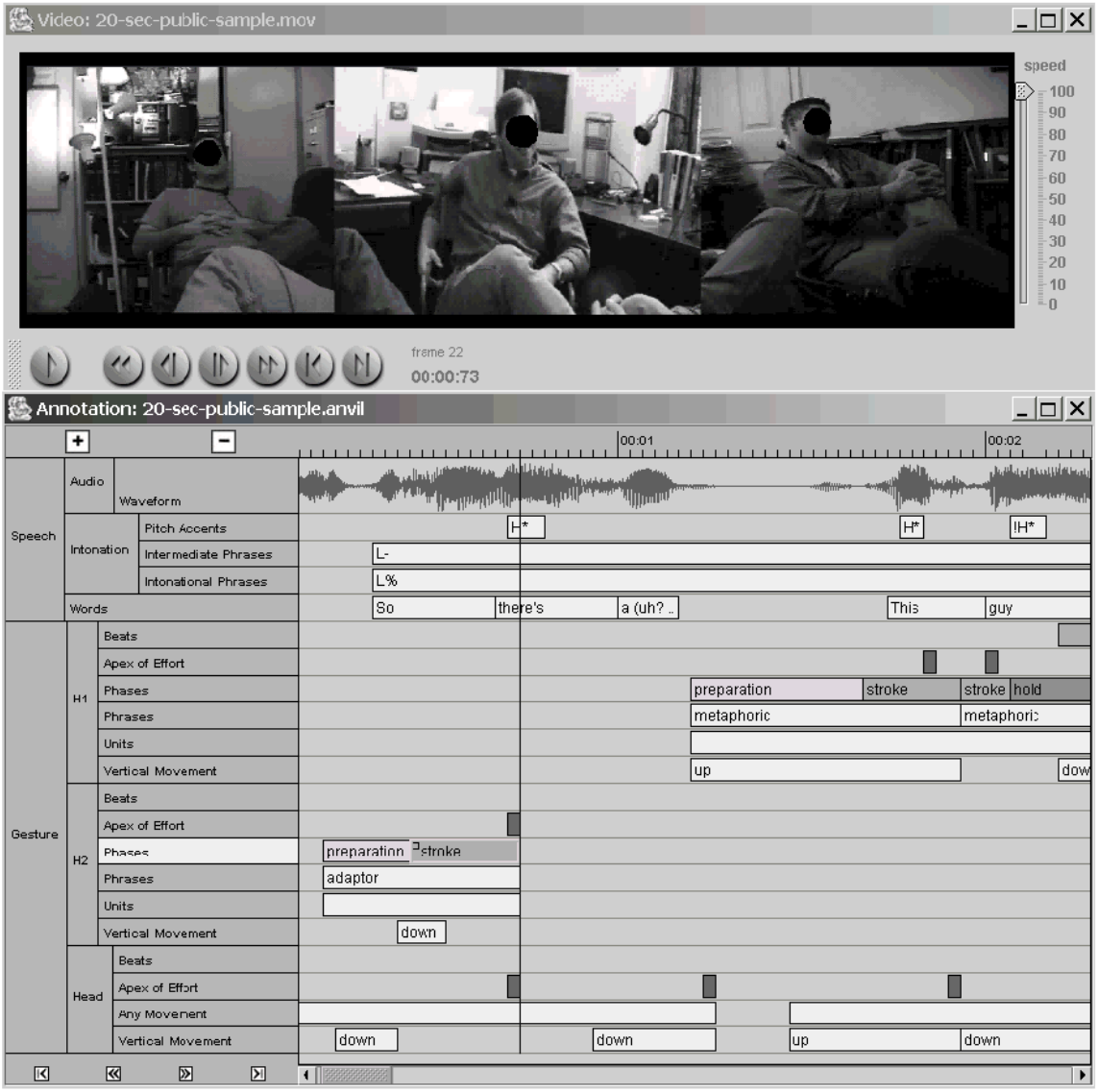
4.4 Post-Annotation Processing

Once the data was annotated, I used two methods of analysis.

The first was observational. Anvil's musical-score layout allows the user to visually notice patterns and alignments. For instance, Figure 6 shows an annotation segment from the "sous-chef" clip. At the word *there's* (under the vertical playback line), one can clearly see that the H* pitch accent, the Hand 2 apex, and the Head apex are aligned. Similarly, under the word *This* to the right of the figure, the H* pitch accent is aligned with the apexes for Hand 1 and Head. The alignments are not exact; the three entities are within three frames of each other, which is just under 100 msec (at 30 frames per second, each frame equals 33 msec). This brings up an important point. The intonational annotations are measurable to less than a millisecond, as Praat allows annotation to that precision. The gesture annotations, however, are limited by the video frame rate. The smallest granularity is one frame, or 33 msec. Therefore, a gesture annotation could be as much as 33 msec off from reality.

Figure 6 looks different from Figure 5, in that the video window contains three video panes, not one. The "sous-chef" clip shown in Figure 6 contains interesting interactional synchrony, and I chose to view all three subjects simultaneously. As described earlier, the three camera shots were spliced together, time-aligned, in a digital video editing tool. The speaker I've annotated is the subject in the middle window.

Figure 6. Sample view of gesture annotation in Anvil, "sous-chef" clip.



The other method of analysis was statistical. Anvil allows all its annotations to be exported with time stamps. The resulting files can be easily imported into statistical packages, which I did for simple analysis. However, such packages can't easily answer questions like "Count all the times a low boundary tone occurs within 300 msec of a downward head movement", or "Produce a table showing how many times each type of tone occurred within 275

msec of each type of movement". Therefore, I wrote computer programming scripts (in the Perl programming language) to do precisely this sort of counting. For example, the latter question is answered in Table 9. (The import of Table 9 will be discussed in subsequent chapters). Appendix C contains the Perl scripts used to produce Table 9.

Table 9

Co-occurrences within 275 msec of movement types and tone types

		Tone Types				
		Pitch Accent	Phrase Accent	Boundary Tone	No Tone	Total
Movement Types	Beat	43	18	4	1	66
	Iconic	71	40	16	0	127
	Metaphoric	117	66	21	0	204
	Deictic	17	17	5	0	39
	Emblem	17	11	4	0	32
	Adaptor	20	13	9	4	46
	No Movement	30	13	7	0	50
	Total	315	178	66	5	564

All told, I analyzed nearly 5,000 video frames and made over 7,500 annotations, which were exported for analysis. The latter figure is based on Anvil's count of 2732 *elements* exported. An element contains at the minimum two annotations: time and type. For instance, a pitch accent element contains the time of occurrence, and the type of pitch accent. Pitch accents and apexes were instants in time, and therefore had only one time point annotated. Most elements were durational, and had two time points annotated (start and stop times), plus the element type, making three annotations. In addition, some elements had attributes; for instance, g-phrases had meaning attributes. Averaging three annotations per element, times 2732 elements, produces my estimate of over 7,500 annotations.

5 Results

As explained in the introduction, I posed five questions, which will be covered in turn below. But first, I'll present more basic findings, which were required before I could look at the original questions. These findings have to do with the relative proximity between annotations for gesture and intonation. As discussed, my methodology has to do with finding gestural and intonational events that occur "near" each other. How near is near? Section 5.1 will discuss that point. The following five sections will then address my five original questions. I'll end the chapter with a section comparing my findings with others reported in the literature.

5.1 Proximity of Gesture and Intonation Annotations

How close in time must two annotations be to be considered "near" each other? As a naïve first estimate, I considered 300 msec. This is relevant phonologically, as words typically average around 300 msec. For instance, the average word length for the "sous-chef" speaker was 303 msec. If, according to McNeill's semantic synchrony rule, gestures align with words carrying the same semantic content, then 300 msec should be close enough to consider gestures and tones near each other.

Yet it would perhaps be more realistic to empirically see how far apart tones and gestures really were in my data, apart from any assumptions. It would also be informative to see how far apart tones occurred from other tones, and how far apart gesture annotations occurred from other gesture annotations.

To this end, I first calculated the average inter-tone interval (ITI) and average inter-gesture interval (IGI) for my data. These intervals are shown in Table 10.

Table 10

Average inter-tone interval (ITI) and average inter-gesture interval (IGI)

Inter-tone interval (ITI)	Inter-gesture interval (IGI)
419 msec	324 msec

The inter-tone interval (ITI) is simply the average of the distances between successive tones. The inter-gesture interval (IGI) requires a bit more explanation. What is typically thought of as a "gesture", following Kendon, is the gesture *phrase* (i.e. a complete gesture). However, as I've explained, I'm investigating the relationship of tones to gestural units on a smaller scale. These include gesture *phases* (the preparations, strokes, holds, and retractions that make up a gesture phrase), and gestural *apexes* (the "peak" of the stroke). As my analyses are based on these smaller units, it is these smaller units that I've included when calculating the IGI. The IGI, therefore, is the average of the distances between successive gesture phases and/or apexes.

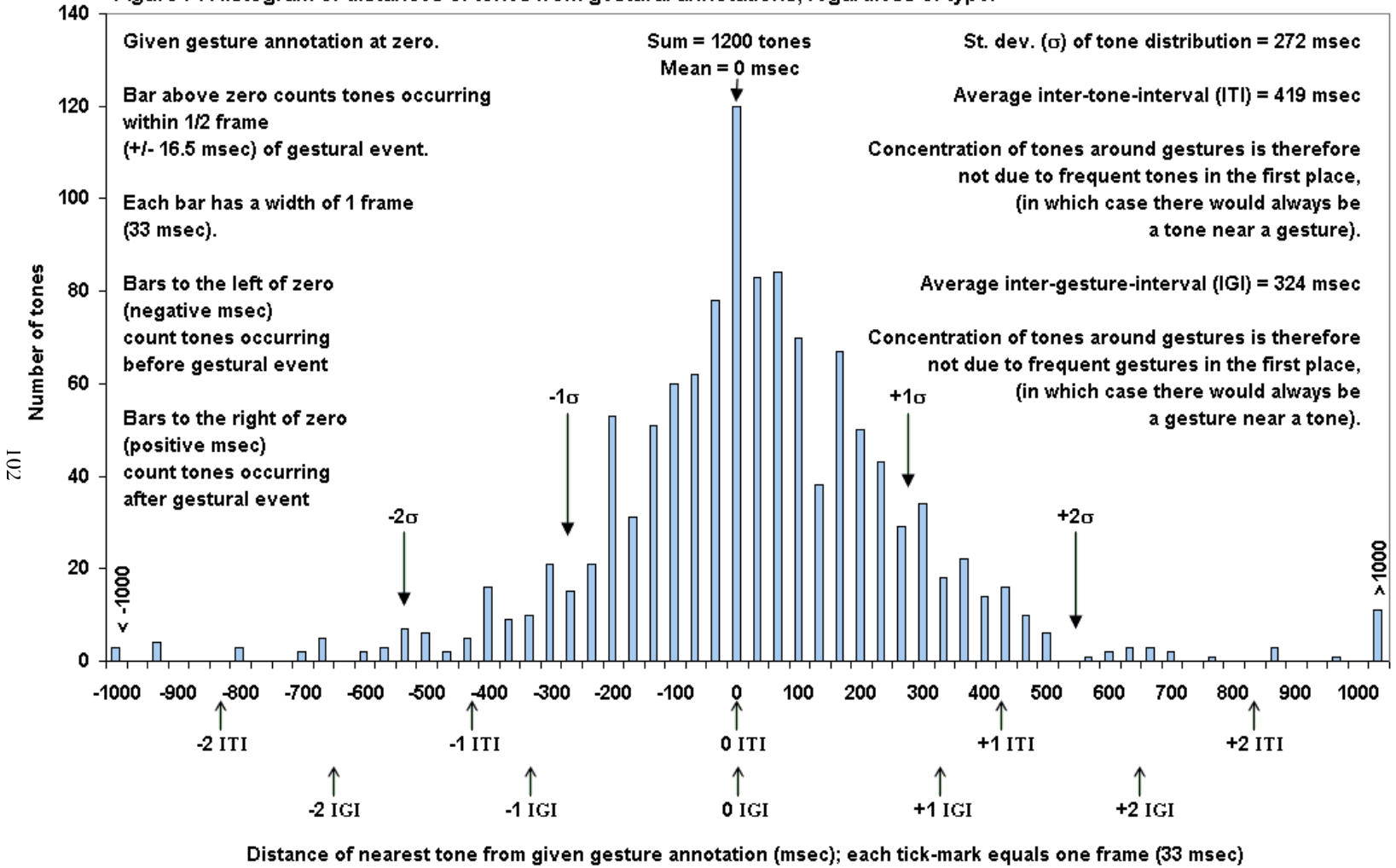
The ITI and IGI were examined in conjunction with the other bit of pre-analysis I did on my data; seeing how far apart tone and gesture annotations really were in my data. To find this out, I performed the following general analysis (using a computer script):

- Take, in turn, every gestural annotation, regardless of type, and regardless of whether the annotation was a starting point of an interval, an ending point of an interval, or an instant in time (e.g. an apex).
- For each gestural annotation, find the nearest tone, regardless of tone type.
- If the tone was to the left of the gesture annotation, count the distance in negative milliseconds. If the tone was to the right, count the distance in positive milliseconds.

- Group all the resulting distances, between all the gesture annotations and their nearest tone, in bins with a width of 33 msec (one frame width, the smallest movement resolution). That is, group all the distances between -0.5 frames and +0.5 frames (-16.5 and +16.5 msec) in one bin, centered on zero. Group all those between +0.5 frames and +1.5 frames (+16.5 and +49.5 msec) in another bin, centered on +1 frame (+33 msec). Group those between -0.5 frames and -1.5 frames (-16.5 and -49.5 msec) in another bin, centered on -1 frame (-33 msec), and so on.
- Count the number of distances in each bin, and plot a histogram. The resulting distribution should give a sense of how, and how far, tones are distributed around gesture annotations, *regardless of type*.

Figure 7 shows the resulting histogram, cumulative for all four data clips.

Figure 7. Histogram of distances of tones from gestural annotations, regardless of type.



In the distribution in Figure 7, the standard deviation is 272 msec. In other words, to answer my pre-analysis question, the majority of the tones in my data, regardless of type, tended to occur within a distance of 272 msec from the nearest gestural annotation.

It may be wondered whether this proximity of tone and gesture annotations is merely due to the fact that tone annotations themselves occur so frequently that there will always be a tone near a gesture annotation. But the ITI is 419 msec, a good deal greater than the standard deviation in Figure 7's distribution. Conversely, one may wonder if gesture annotations are so frequent that there will always be a gesture annotation near a tone. But the IGI of 324 msec is again greater than the standard deviation. In other words, the tones are clustering around gesture annotations for reasons other than a possible high frequency of either type.

In fact, the clustering in Figure 7 is worth discussing on its own, apart from answering our initial question of how far apart tones and gestures really occur in the data. The mean of the distribution of 1200 tones is exactly 0 msec. Certainly randomly-distributed tones would also average to zero, when comparing their distances from a given gesture event. But the histogram curve of randomly-distributed tones would be flat, not peaked near the center as in Figure 7. The peak means that tones tend to occur near gestural events.

One might also criticize Figure 7 as being overly general, as it combines data from all four speakers. But the distribution means for the individual speakers are 0, -21, +36, and -16 msec, respectively, for the "cupboards", "musicians", "drywall", and "sous-chef" subjects—all averaging around a frame or less, which is the minimum resolution possible for movement analysis. Thus, regardless of tone type, and regardless of gestural event, and regardless of the individual speaker in my data, the events of gesture and intonation tended to cluster, on average, to within a few milliseconds of each other, which is remarkable.

To summarize this pre-analysis, the goal was to determine how close a tone and a gesture annotation needed to be to be considered "near" each other. My initial naïve window was 300 msec, based on word-length. The subsequent empirical analysis showed that looking 272 msec away from a given gesture annotation would be far enough to find the majority of tones likely to occur nearby (as 272 msec is the standard deviation of the distribution of tones around gesture annotations). The two figures, one phonologically motivated, and the other based on an overview of the data, are similar. To be conservative, I chose the smaller of the two windows, 272 msec, as my window of analysis. To avoid having to explain this unusual number each time I present a piece of analysis, I chose to round it up to 275 msec.

Therefore, in the analyses which follow, all calculations are based on a co-occurrence window of 275 msec.

5.2 Does Bolinger's Parallel Hypothesis Hold?

As discussed, Bolinger claimed that pitch and body parts move up and down in parallel, reflecting increased or decreased tension. As also discussed, I specifically annotated vertical movement of the head and hands, alongside intonation annotations. How do these relate, if at all? Table 11 shows the co-occurrences of vertical movement and tones within 275 msec.

Table 11

Co-occurrences within 275 msec of vertical movement and tone types

	H*	!H*	L*	L*+H	L*+!H	L+H*	L+!H*	H+!H*	X*	H-	!H-	L-	X-	H%	L%	X%	No tone	Tone totals
Hand(s) up	47	17	0	0	0	12	0	0	41	2	0	39	11	1	19	1	11	201
Hand(s) down	62	23	2	0	0	16	0	0	38	0	0	49	7	0	20	1	14	232
Head up	54	17	1	0	0	11	0	0	37	2	0	59	9	0	35	4	18	247
Head down	64	15	2	0	0	14	0	0	48	1	0	59	8	1	27	2	18	259
No vertical movement	24	5	1	0	0	4	0	0	4	1	0	20	7	1	4	1	0	72
Movement totals	251	77	6	0	0	57	0	0	168	6	0	226	42	3	105	9	61	1011

If movement and pitch were to move in parallel, then patterns should emerge in Table 11. For instance, look at the rows for *Hand(s) up* and *Hand(s) down*. Looking across these rows, are there any tones for which *Hand(s) up* and *Hand(s) down* pattern very differently? No. For instance, the very first tone, *H**, occurred with a *Hand(s)* up 47 times, and with a *Hand(s) down* 62 times. *H** is a high tone, and we might expect it to occur more often with hands moving up than down. But the opposite is the case, although 62 is not significantly greater than 47, compared to the average of the two (which would be the case if the events were randomly distributed). Table 12 shows this distribution.

Table 12

Distribution of an H tone within 275 msec of vertical hand movement*

		Distribution with H*	
		Observed distribution with H*	Expected random distribution with H*
Vertical Hand Movement	Hand(s) up	47	54.5
	Hand(s) down	62	54.5

As mentioned, an H* tone was observed with hand(s) up 47 times, and with hand(s) down 62 times, for a total of 109 observations. If those observations were randomly distributed between the two cases, they would occur 54.5 times with each. The distribution in Table 12 is not significant on a chi-squared test[15], meaning that H* does not pattern significantly according to vertical hand movement. As also mentioned, the trend (though not significant) is opposite of what Bolinger would predict: we have the hand lowering on a high tone.

In the column for *L-,* a low phrase tone, *Hand(s) down* fares slightly better than *Hand(s) up*, which would make sense according to Bolinger's hypothesis. But the difference is only 49 to 39, which is also not significant, as shown in Table 13.

[15] I have applied Yate's correction to all 2x2 chi-squared tables in my dissertation.

Table 13

Distribution of an L- tone within 275 msec of vertical hand movement

		Distribution with L-	
		Observed distribution with L-	Expected random distribution with L-
Vertical Hand Movement	Hand(s) up	39	44
	Hand(s) down	49	44

Looking at the rows for *Head up* and *Head down* is similarly fruitless. It might be expected that the head would move up or down based on tones at the *ends* of sentences, as Bolinger specifically mentions this phenomenon. Tones at the end of sentences would be boundary tones, as shown in the columns for *H%* and *L%*. Yet the low boundary tone *L%* occurred more often with the head moving up (35 times) than with the head moving down (27 times), although again the difference is not significant, as shown in Table 14.

Table 14

Distribution of an L% tone within 275 msec of vertical head movement

		Distribution with L%	
		Observed distribution with L%	Expected random distribution with L%
Vertical Head Movement	Head up	35	31
	Head down	27	31

I should explain that Table 11, in its entirety, is not suitable for chi-squared analysis, because the occurrences in the cells are not all independent. Pitch accents can occur closely enough to edge tones to be counted twice in the table. The same is true for phrase tones and boundary tones, which often occur together at the ends of intonational phrases. Additionally, a tone might coincide with vertical movement of both the hands and the head, allowing it to be counted twice. Therefore, the table as a whole may only be used for detecting general patterns. However, certain subsections of the table do contain independent data. The sections I've discussed in the previous paragraphs contain independent cells. For example, comparing up or down head movement, versus a high or low boundary tone, would entail independent data. This is because the head can't move in both directions at once, and the larynx can't produce a high and low tone simultaneously.

Therefore, according to my data, Bolinger's parallel hypothesis has no support.

In fairness, it may be argued that for some of ToBI's tones, it might not be clear whether the tone is "up" or "down" in the first place. This is certainly true of bitonal pitch accents, which by definition include both a high and a low component. And even a simple high pitch accent, H*, is a local *high,* surrounded by areas of *low*. However, edge tones—the phrase tones (H- and L-) and the boundary tones (H% and L%)—are typically moving in a single direction. Yet none of these pattern in parallel with vertical movement.

One reason Bolinger's prediction was not borne out in my data may have to do with an exception Bolinger himself pointed out. As discussed earlier, Bolinger wrote (1986, p. 200):

> One frequent exception … is the use of a downward thrust of the jaw to mark an accent … Since intonation marks accents much of the time with a rise-fall, pitch and head movement may go in opposite directions…

I found this exception numerous times in my data, where a high pitch accent (H*) coincided with a downward head nod. What's more, the head nod was typically followed by a slight "bounce" back up of the head. It was often the case that the nucleus was an H*, and that the nucleus was near the end of the sentence. This resulted in the head "bouncing" up right at the end of the sentence, which is the opposite of Bolinger's prediction.

Another problem with proving Bolinger's hypothesis may have been an artifact of the ToBI annotation guidelines. As mentioned, phrases (both intermediate and intonational) are to be aligned with word boundaries. This means that the phrase ends when the phrase-final word ends. Often, in a final word, voicing would end early, but the final unvoiced segment would end much later. For example, in the phrase-final "*the hinges are different*" from the "cupboards" clip, the word *different* lasts 500 msec, but voicing ends after the H* nucleus on the syllable *dif*, 400 msec before the end of the word. In such cases, I would duly record the final tone at the end of the word, well after voicing had finished. This by itself seems to be a slightly troublesome issue, internal to ToBI. But it also made correlation with body movement potentially problematic. Movements which appeared to be correlated with speech sometimes seemed to end early in the final word, perhaps correlating with the end of voicing, not the end of the word. Thus, potential alignments of head lowering with low boundary tones might be missed statistically, if the speaker were aligning head movements with the end of voicing, instead of with the end of the word (where I placed the tone, per ToBI's instructions). In scanning the data by eye, I noticed only a few places where this seemed to be the case, but the issue is worth mentioning.

A third problem with proving Bolinger's hypothesis may be that Bolinger included facial movements in his examples, which I didn't include in my annotations.

Thus, like McClave, I found no empirical support for Bolinger's hypothesis. The only relationship between intonation and vertical movement I could detect was downward head nods

on pitch accents, marking stress. If the nucleus was near the end of a sentence, and if the sentence ended in a low pitch, then yes, the head did move down towards the end of a falling-tone sentence, but only because of a nod on the nucleus. Perhaps Bolinger's claim has some merit. But it may be limited to certain situations, of the kind he describes in his examples, and which did not occur in my data.

5.3 How Do Gesture and Intonation Unit Boundaries Align?

I checked for alignments between the three levels of intonational units, and the four levels of gestural units. The intonational units were, from smallest to largest, pitch accents, intermediate phrases, and intonational phrases. The gestural units, again from smallest to largest, were apexes of strokes, gestural phases (especially strokes), gestural phrases, and gestural units. Of all these possible combinations, I found two pairs that aligned. Apexes aligned with pitch accents, and gestural phrases aligned with intermediate phrases. I'll discuss each in turn.

5.3.1 Apexes of Strokes Align with Pitch Accents

It's very clear that apexes align with pitch accents in my data. This is not an entirely new finding; numerous researchers have observed that gestural strokes line up with stressed syllables. My finding uses slightly different units: the apexes of gestural strokes (instead of the strokes themselves), and pitch accents (instead of stressed syllables). I also lend empirical weight to this common observation. Table 15 shows how pitch accents align with apexes. For simplicity, I have collapsed all other types of each modality into an "other" category. I have also ensured that Table 15 contains independent data, by removing any annotations in my data that could occur in two table cells (e.g. a phrase tone and a boundary tone within 275 msec of each other).

Table 15

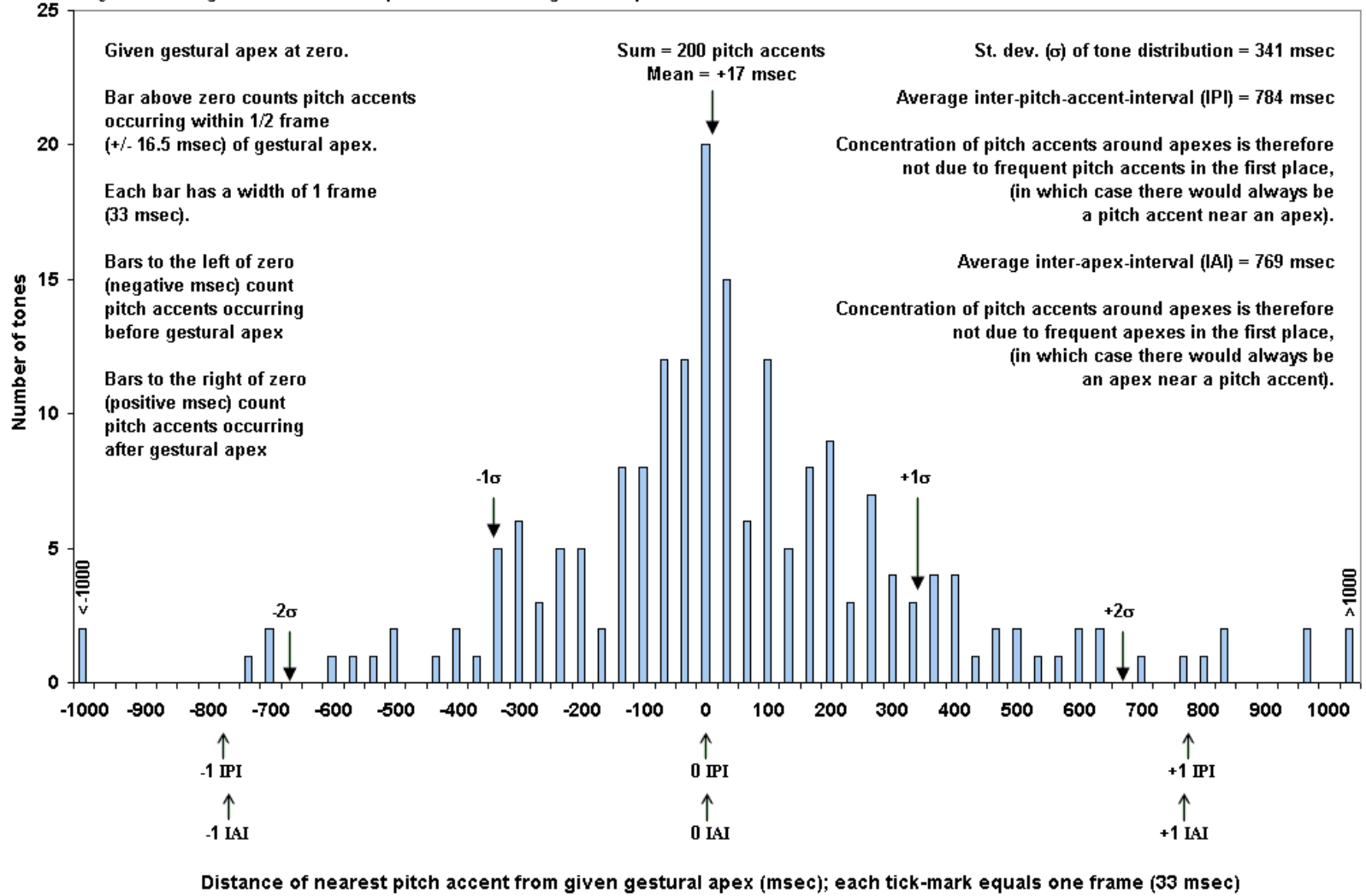
Co-occurrences within 275 msec of pitch accents and apexes

		Tones	
		Pitch Accent	Other (non-pitch accent)
Movement	Apex	83	28
	Other (non-apex)	43	46

The distribution in Table 15 is very highly significant on a chi-squared test ($p < .001$). Although pitch accents certainly occur with non-apexes (often because there is speech without gesture), and although apexes sometimes occur with non-pitch accents, the significant tendency is for them to occur together.

The relationship between apexes and pitch accents can be seen another way statistically. Recall that Figure 7 plotted a histogram of the distribution of any tone around any gestural annotation. As explained, this is really a very general comparison, with no hypothesis driving it (other than the general relationship of gesture and speech). But we can test a more specific hypothesis with this type of calculation: the distribution of pitch accents around apexes. I used the same methodology as I used for Figure 7, except that I only looked at apexes, and then only found the nearest pitch accent to each apex. The resulting histogram is shown in Figure 8.

Figure 8. Histogram of distances of pitch accents from gestural apexes.



As in the general distribution, the distribution of pitch accents relative to apexes is centered very close to zero; the mean is +17 msec, or half a frame. The means of the individual speakers is +50, +10, +48, and -92 msec, respectively, for the "cupboards", "musicians", "drywall", and "sous-chef" subjects. So the means of the individuals show greater variation from zero than they did in the general case, but are still within a few frames of zero, which is quite short phonologically (much smaller than most syllables).

There is one other difference between Figures 7 and 8. The distribution in Figure 8 is not as regular as the distribution in Figure 7, perhaps because there are fewer data points counted (200 pitch accents, compared to 1200 tones in all).

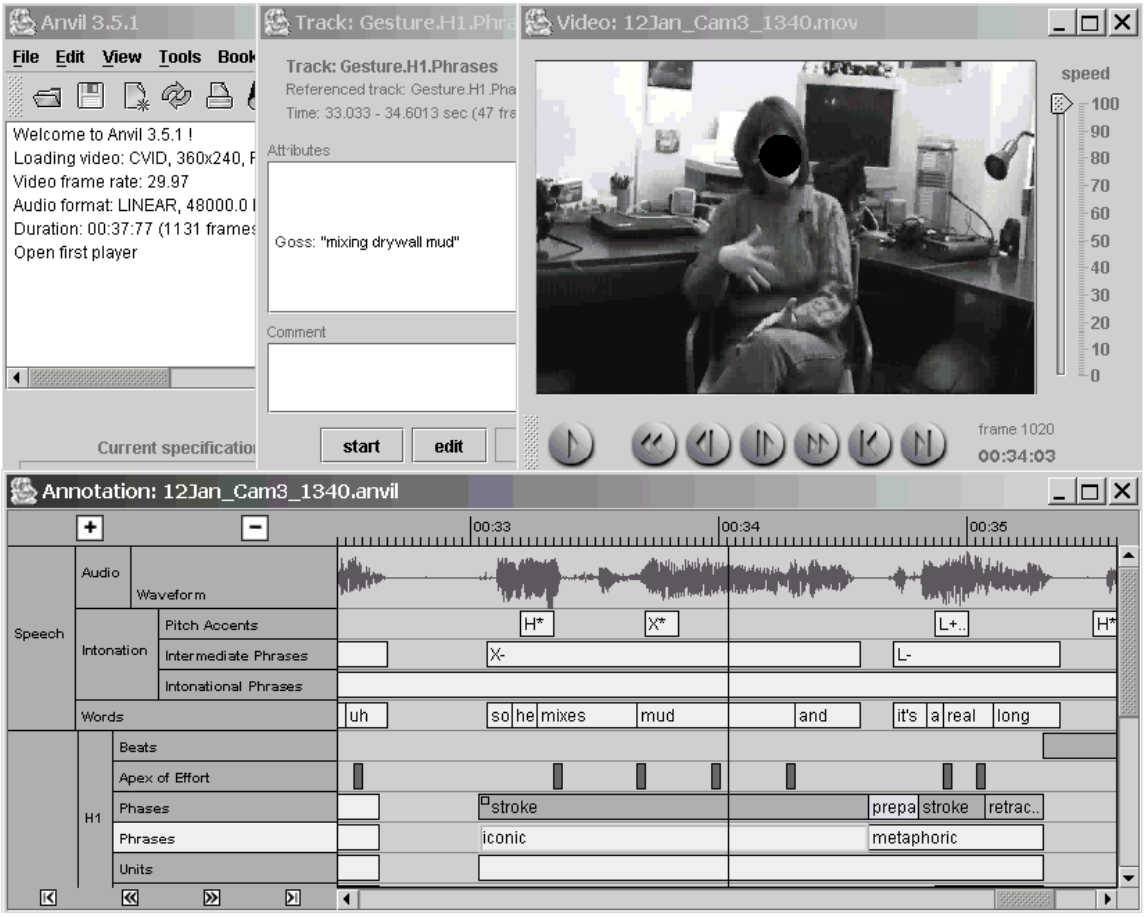
The tight relationship between apexes and pitch accents can also be clearly seen observationally in Anvil's timeline, as was shown in Figure 6 above. The two phenomena co-occur repeatedly and obviously in a scan of Anvil's musical-score layout.

5.3.2 Gestural Phrases Align with Intermediate Phrases

In addition to finding a correlation between apexes and pitch accents, I also discovered a correlation between gestural phrases, or g-phrases, and intermediate phrases. The boundaries of g-phrases aligned with the boundaries of intermediate phrases. Typically, a g-phrase was found to align with a single intermediate phrase. Often, multiple g-phrases (never more than three) were found within a single intermediate phrase. On these occasions, the boundaries of the g-phrases occurred at places within the intermediate phrase where there was a syntactic constituent boundary, or a slight pause not deserving of an intermediate phrase boundary, or both.

Over two-thirds of the g-phrases in my data aligned clearly with an intermediate phrase. Figure 9 shows an example, from the "drywall" clip.

Figure 9. Example of gestural phrases aligning with intermediate phrases.



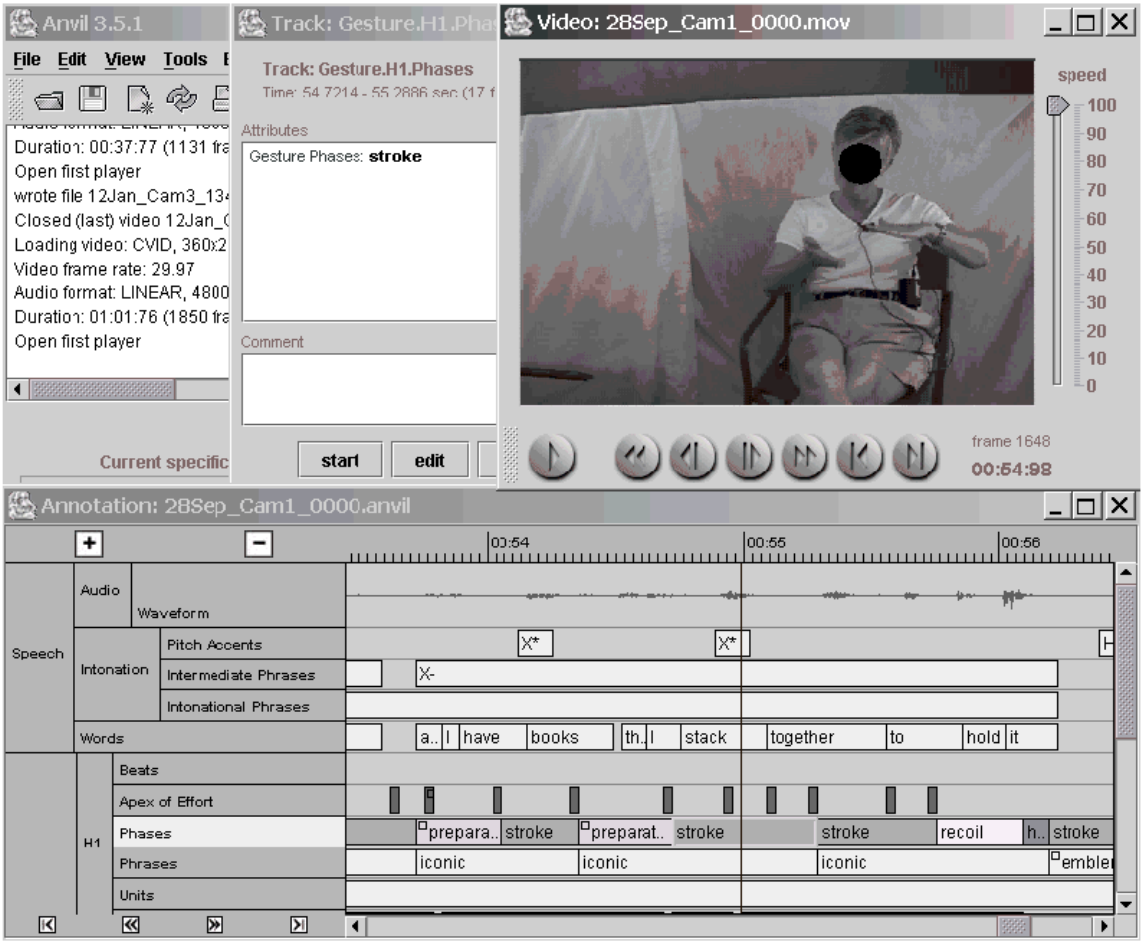
In Figure 9, the first intermediate phrase shown comprises the words "*so he mixes mud and*". Below that on the annotation board, one can see a corresponding g-phrase, an iconic in which the speaker waves her hand around as if mixing something (a snapshot is shown in the video window). The next intermediate phrase, comprising "*it's a real long*", is matched by a metaphoric g-phrase. (Although not shown in Figure 9, in this metaphoric the speaker waves her hands dismissively.)

Note that both g-phrases in Figure 9 slightly precede their corresponding intermediate phrases. The typical pattern, for those g-phrases that clearly aligned with an intermediate phrase,

was for the g-phrase to start and end slightly ahead of its corresponding intermediate phrase. An approximate average for this lead time was 3 frames (100 msec), for both the start and end of the respective phrases. However, there was quite some variation in timing. Some g-phrase boundaries (both the start and end boundaries) occurred as much as 10 frames (330 msec) before, or as much as 4 frames (133 msec) after, their corresponding intermediate phrase boundary.

As mentioned, multiple g-phrases often occurred within a single intermediate phrase. Figure 10 shows an example of this, from the "cupboards" clip.

Figure 10. Example of multiple gestural phrases aligning with a single intermediate phrase.



In Figure 10, there are three successive iconic g-phrases, in which the speaker depicts stacking books higher and higher with each g-phrase. (The handshape changes between the g-phrases; otherwise, they might be considered a single multi-stroke g-phrase). The three g-phrases fit within a single intermediate phrase, although the g-phrase boundaries can be seen to coincide with the syntactic boundaries in "*and I have books* | *that I stack together* | *to hold it*" (where I've marked syntactic boundaries in the preceding italics with vertical bars). Note that the starts of the second and third g-phrases precede the syntactic boundaries I've mentioned, although the heart of the gestures, the apexes of the strokes, don't occur until the appropriate syntactic constituents. There is also a slight pause between *books* and *that*, although the pause did not merit a intermediate phrase boundary, based on the intonational tune[16].

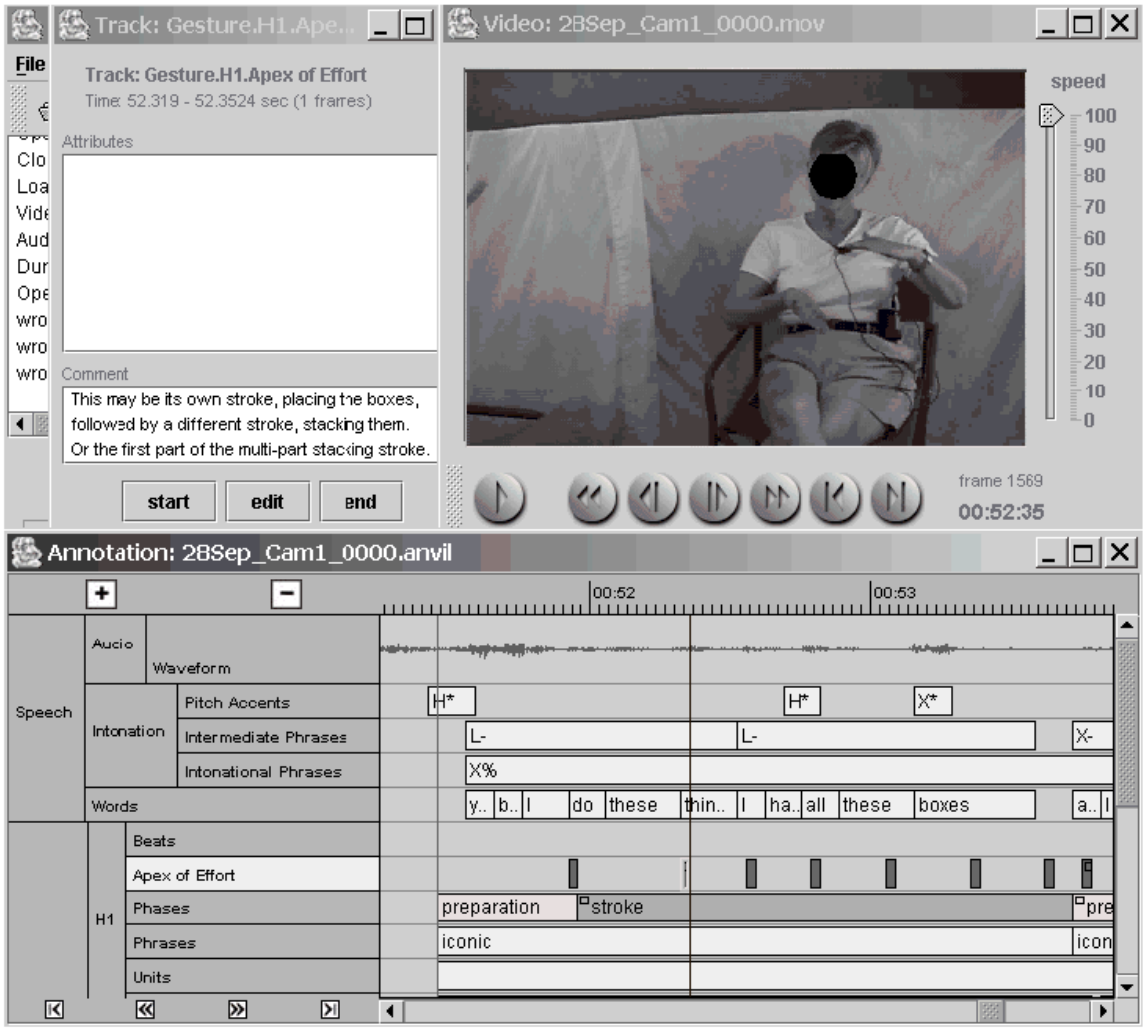
Typically, an entire g-phrase aligned with an intermediate phrase. Occasionally, however, it was clear that a g-phrase aligned with an intermediate phrase only when disregarding post-stroke holds, retractions, or recoils within the g-phrase. These internal components are included within g-phrases by definition, following Kendon's hierarchical packaging. However, there may be some different quality about these post-stroke components. Occurring after the heart of the gesture, they may have a less important status in terms of timing with speech.

I've said that g-phrase boundaries align with intermediate phrase boundaries, and that multiple g-phrases may exist within a single intermediate phrase. Is the converse true? Do multiple intermediate phrases ever occur within a single g-phrase? There were a few instances where a g-phrase appeared to span several intermediate phrases. Yet, upon reviewing my annotations, in every such instance there was some unusual point within the g-phrase which I had deliberated over or commented on during initial coding. This markedness may have been a

[16] This pause corresponds to a level 2 break index in the ToBI scheme, which is designed to include exactly such cases, where pausing indicates a break but intonation doesn't.

noticeable change in direction in what I initially labelled a multi-part stroke, but which could have been two strokes (and hence two g-phrases). It may have been a post-stroke hold with an inertial beat, which could also have been labelled a separate stroke. Or it may have been an especially long and distinctive preparation which might deserve to be its own stroke. Most importantly, this unusual point in the g-phrase was always near an intermediate phrase boundary. As an example, see Figure 11, which takes place just prior to the data clip in Figure 10.

Figure 11. Example of a single gestural phrase possibly spanning multiple intermediate phrases.

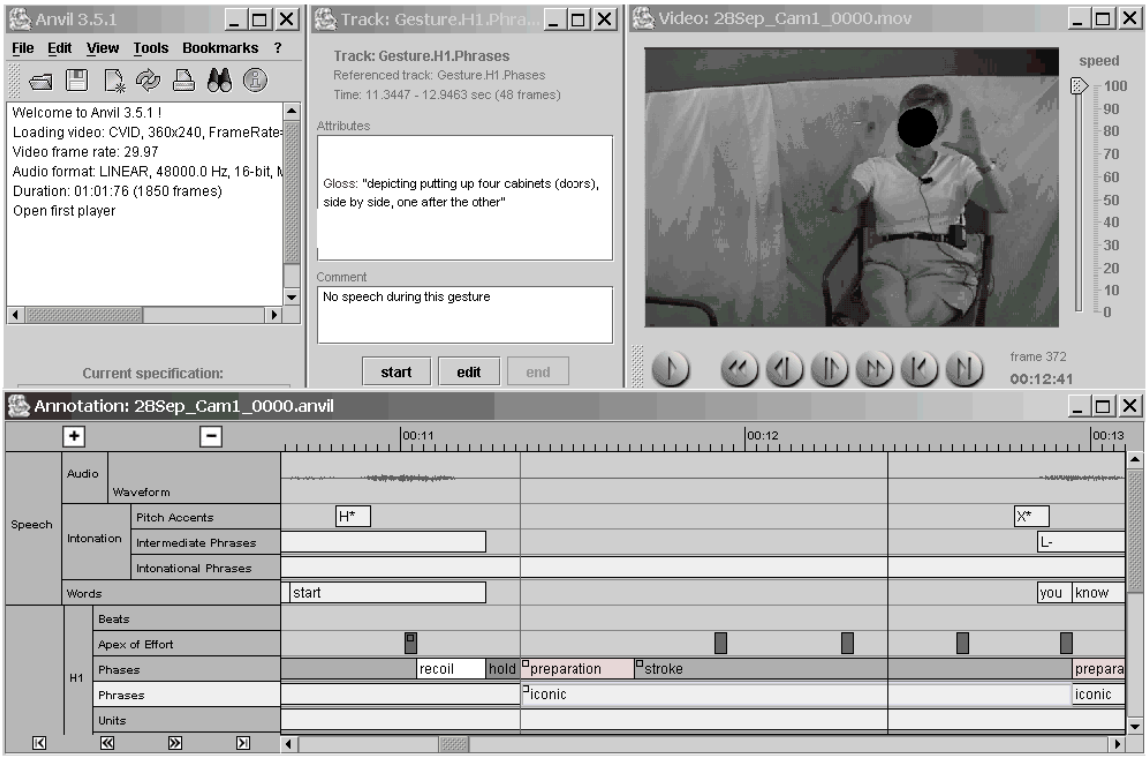


In Figure 11, there is a single iconic g-phrase spanning two intermediate phrases (covering "*yeah but I do these things*" and "*I have all these boxes*"). But at a point just prior to the intermediate phrase boundary, there is a change of character in the stroke (as noted in the annotation comment in the upper left), which may indicate that there could be two strokes, and therefore two g-phrases, one for each intermediate phrase.

Thus, although there are a few cases where a g-phrase spanned multiple intermediate phrases, in every one of those few cases I initially considered alternate labelling which would have resulted in multiple g-phrases, each aligning with a single intermediate phrase. Therefore, it may well be the case that multiple intermediate phrases do not occur within a single g-phrase, although my data is too inconclusive to be sure.

There were three other interesting points about g-phrases and intermediate phrases. First, during disfluencies of speech, aborted g-phrases aligned with aborted intermediate phrases. Second, g-phrases did not align directly with the larger intonational phrases (although they did so indirectly, in that intonational phrases contain intermediate phrases). Third, though g-phrases may align with intermediate phrases, the converse need not be true. Intermediate phrases need not align with g-phrases, by the mere fact that people often speak without gesturing, but rarely gesture without speaking. But several times, subjects did in fact gesture during short speech pauses. These gestures, while certainly supporting the surrounding speech, seemed to convey an idea by themselves during the speech pause. The interesting point relevant to my discussion is that in these cases, the g-phrase fit neatly within the speech pause. Figure 12 shows an example of this, from the "cupboards" clip.

Figure 12. Example of gestural phrase fitting within a speech pause.



In Figure 12, the speaker is in the middle of saying "*and so I start <1.6 second pause> you know putting them up and everything*". During the pause, she makes an iconic gesture, in which she places four cupboard doors in the air, one after the other, side by side. It's as if she demonstrates her cupboard-installation activities first silently, then discusses them with speech afterwards.

In the previous subsection, I noted that gestural apexes consistently aligned with pitch accents. I then showed this visually, with a tightly clustered histogram of their time differences (Figure 8). In this subsection, I've discussed how g-phrases typically, but not always, align with intermediate phrases. When they do align, it is sometimes with adjustments, such as disregarding a hold, or aligning with a syntactic boundary. In other words, the correlation between g-phrases and intermediate phrases is looser than the correlation between apexes and pitch accents. This

looser correlation is also revealed in histograms. Figure 13 shows the distribution of the starts of g-phrases with the starts of intermediate phrases. As in the earlier histograms, Figure 13 was generated by naïvely looking for the nearest intermediate phrase start from a given g-phrase start, no matter how far away, regardless of whether or not it seemed apparent that the g-phrase and intermediate phrase were meant to align. Figure 14 shows the distribution of the ends of g-phrases with the ends of intermediate phrases, generated with the same naïve strategy.

Figure 13. Histogram of distances of intermediate phrase starts from gestural phrase starts.

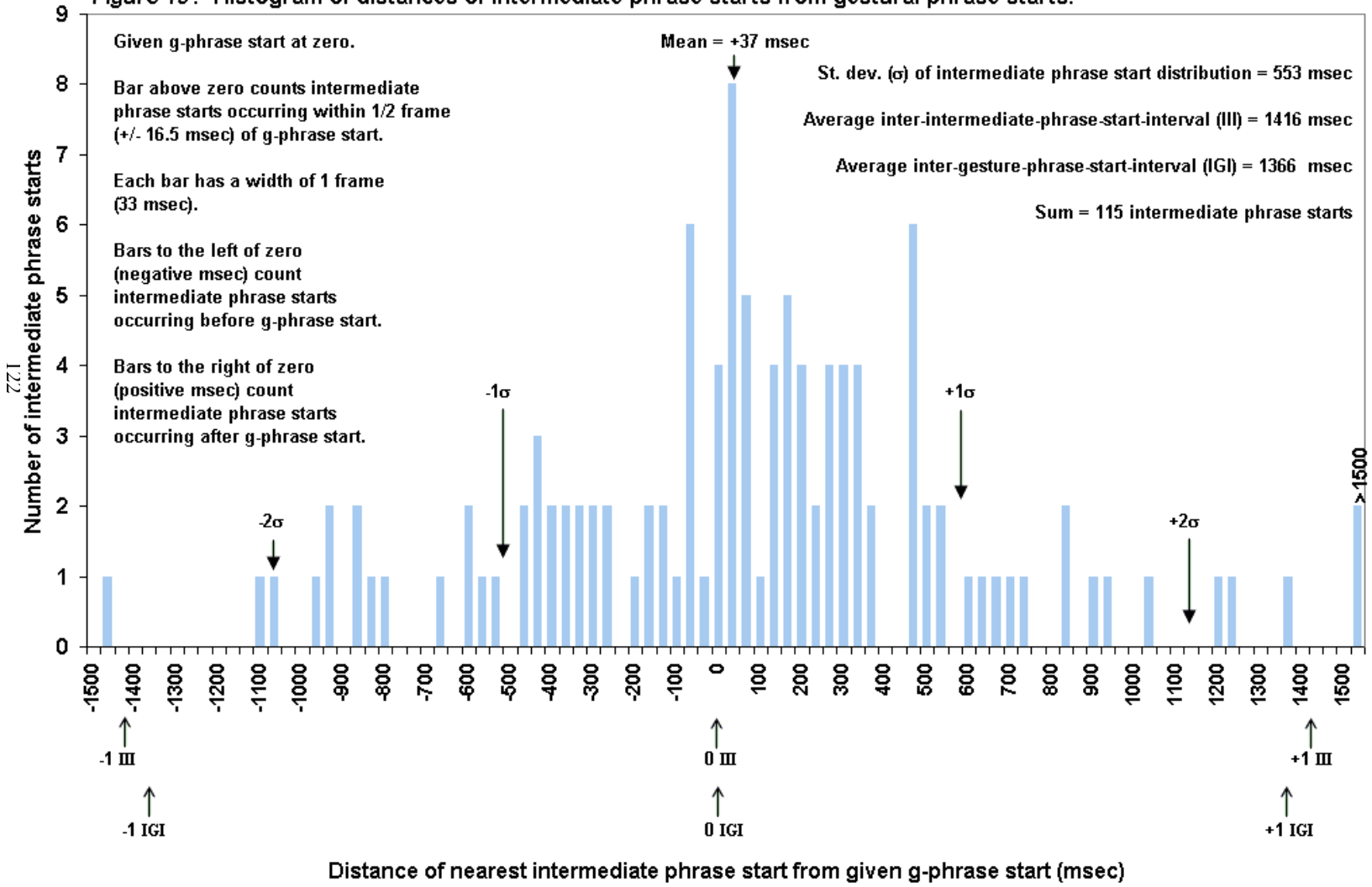
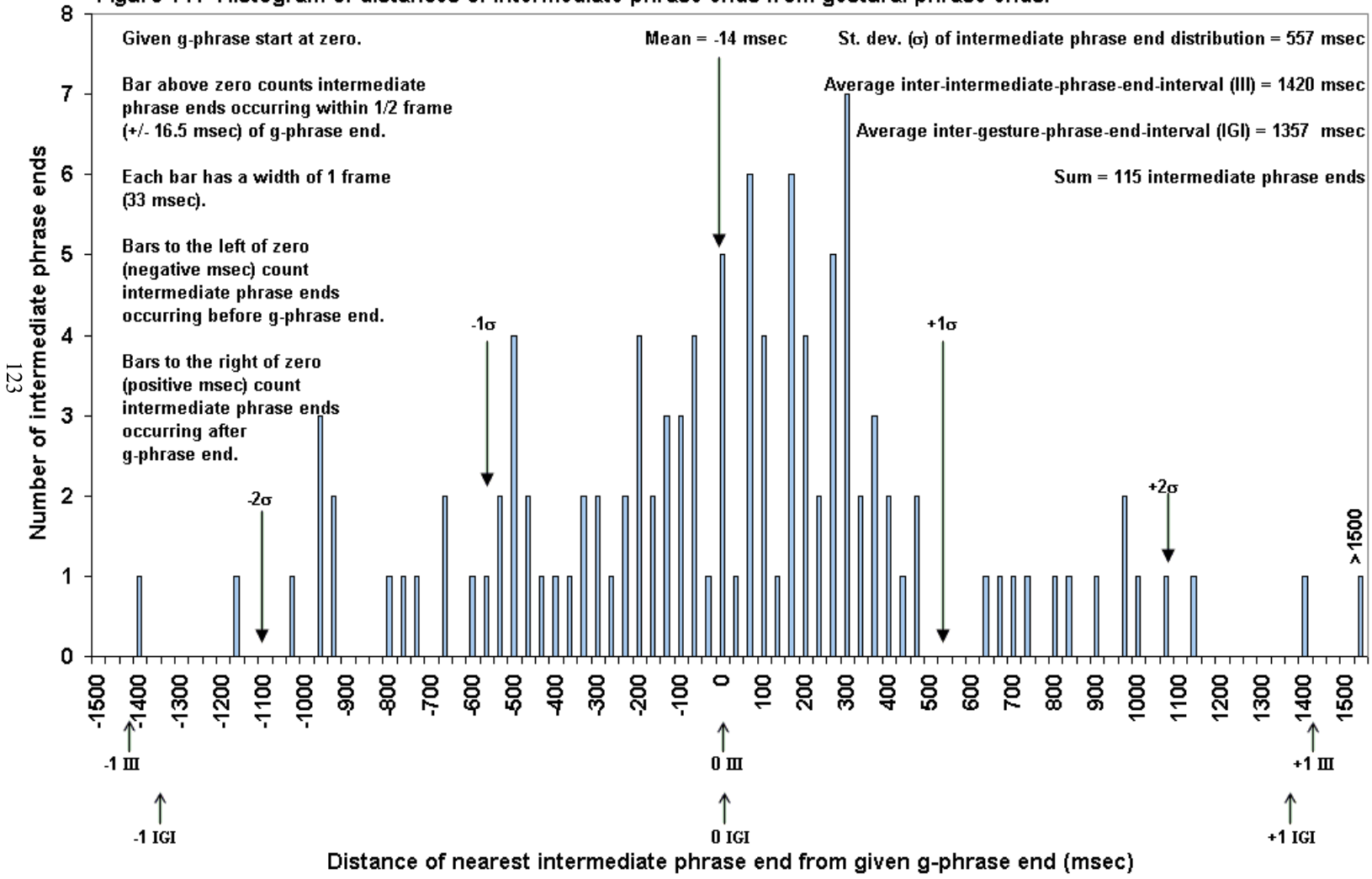
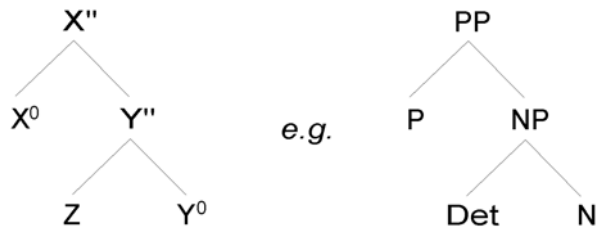


Figure 14. Histogram of distances of intermediate phrase ends from gestural phrase ends.

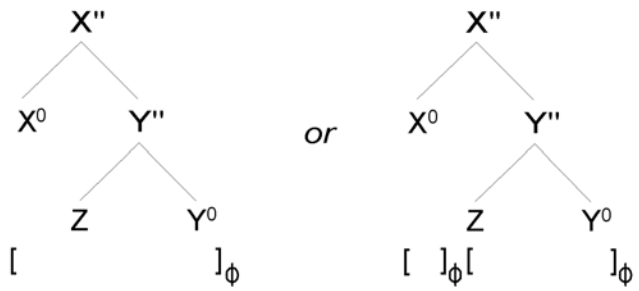


As can be seen in Figures 13 and 14, when compared to the clustering of apexes and pitch accents, the starts and ends of g-phrases don't cluster as tightly with the starts and ends of intermediate phrases, respectively. The distributions do peak somewhat towards the middle, and do have averages quite close to zero (+37 msec for starts, -14 msec for ends). But the distributions are not as uniform, and the standard deviations are larger (553 and 557 msec).

I have argued that g-phrases align with intermediate phrases, but not perfectly. Could it be the case that g-phrases are actually aligning with phonological phrases? I have been looking at Pierrehumbert's intonational hierarchy, but there is also the prosodic hierarchy. In the prosodic hierarchy, the phonological phrase is the next level smaller than an intonational phrase, so perhaps the phonological phrase is a suitable correlate for the g-phrase. Inkelas and Zec (1995) defined a phonological phrase as a syntactic X", shown below on the left. An example of X" is a prepositional phrase (PP), as shown below on the right.



Depending on the circumstances (and on the particular theoretician), the phonological phrase (denoted $[\ldots]_{\Phi}$) could either map to the entire X", or it could map separately to both the head phrase and the complement, as shown below.



How well does the phonological phrase match up to g-phrases in my data? There are certainly some good matches. For example, in the "drywall" clip, the PP *in my neighborhood* is accompanied by a g-phrase. This PP is clearly a phonological phrase, so the match is a good one. However, there are a number of g-phrases which accompany syntactic constituents much larger than prosodic phrases. The "drywall" subject provided three such examples: (1) *it's neat to see the process, though*; (2) *like a new configuration of one of her rooms*; and (3) *and I watched this from day to day*. All three of these stretches of speech were accompanied by both a g-phrase and an intermediate phrase. The prosodic phrase is too small a chunk for these cases. Thus, although the prosodic phrase is sometimes a good match for a g-phrase, the intermediate phrase provides a better overall match, handling the longer g-phrases.

To summarize this subsection, gestural phrases patterned with intermediate phrases in my data. One or more g-phrases typically occurred with a single intermediate phrase. If there were multiple g-phrases in an intermediate phrase, the g-phrase boundaries occurred at syntactic and/or pausal boundaries within the intermediate phrase. G-phrases typically slightly preceded their counterpart intermediate phrases, and g-phrases sometimes fit neatly within pauses in the speech. I'll discuss the theoretical implications of these findings in the next chapter.

5.4 How Do Gesture and Intonation Unit Types Correlate?

In the previous section, I looked for temporal alignments between the various hierarchical levels of intonational and gestural units. The question in the present section is: Are there correlations between the various types of each modality? As explained above, gestures come in four flavors: beats (rhythmic emphasis), iconics (pictorial), metaphorics (abstract), and deictics (pointing). In addition, there are emblems (e.g. the "OK" sign), and adaptors (e.g. scratching oneself). Do any of these types pattern with the various intonational types (pitch accents, intermediate phrases, and boundary tones)?

The answer is no. Table 16 shows the distribution of movement types and tone types within 275 msec of each other.

Table 16

Co-occurrences within 275 msec of movement types and tone types

		Tone Types				
		Pitch Accent	Phrase Accent	Boundary Tone	No Tone	Total
Movement Types	Beat	43	18	4	1	66
	Iconic	71	40	16	0	127
	Metaphoric	117	66	21	0	204
	Deictic	17	17	5	0	39
	Emblem	17	11	4	0	32
	Adaptor	20	13	9	4	46
	No Movement	30	13	7	0	50
	Total	315	178	66	5	564

Table 16, like Table 11 shown earlier, is not suitable for chi-squared analysis, because the data are not completely independent. This is because tones can occur closely enough to each other to be counted twice. Likewise some movements can be superimposed upon each other and counted twice. But Table 16 is suitable for looking at general trends. When looking across a row or down a column, are there any types that pattern unusually? No. The co-occurrences are distributed pretty much as one would expect them to be, based on the relative proportion of each type. For instance, there are more pitch accents than any other tone, and every movement type has more co-occurrences with pitch accents than with the other tones (except deictics; see below). Similarly, there are more metaphorics than any other movement type, and every tone has more co-occurrences with metaphorics than with any other movement type.

There are a few rows that hint at patterning differently. As mentioned, deictics are equally distributed between pitch accents and phrase accents (17 co-occurrences each), even though there are many more pitch accents than phrase tones in the data. But when looking at deictics versus other types of movement (after removing duplicate entries from the data to allow a

chi-squared test), deictics did not pattern significantly differently than other movement types. One problem here may be that there are too few deictics in my data to permit proper significance testing. The same may be true of emblems and adaptors.

Thus, I found no correlation between movement types and intonation types. Upon reflection, it might be surprising if I had. I can think of no theoretical motivation why, for example, phrase tones would be more likely to occur with metaphorics than with iconics. Moreover, Susan Duncan has emphasized that the idea of a gesture belonging to only one of the four categories is a convenient fiction, and that in reality gestures are typically composites of multiple types (personal communication, August 26, 2002). However, since I was easily able to check for such correlations once the data was annotated, it was interesting to investigate. The results were negative.

5.5 How Do Gesture and Intonation Meanings Correlate?

McNeill's semantic and pragmatic synchrony rules state that co-occurring gesture and speech present the same semantic information, or perform the same pragmatic function. By speech, McNeill meant the actual words spoken. Is the same true for intonational meaning? Does intonational meaning correlate with gestural meaning, either semantically or pragmatically?

The meaning carried by English intonation is largely pragmatic, as opposed to semantic. Therefore, the semantic synchrony rule would not apply to intonation. But I found that the pragmatic synchrony rule does.

Pragmatic synchrony between intonation and movement doesn't occur with every gesture, unlike the meaning correlation between words and gesture. Rather, it surfaces only in occasional (but striking) situations. This is hardly surprising. Intonation has only a limited set of pragmatic functions, so it couldn't synchronize pragmatically with gesture in every phrase. In contrast, words and gesture each have nearly limitless expressive potential.

There are hundreds of thousands of words in the English language, which can be combined syntactically in infinite ways to describe possibly any human thought imaginable. Intonation, on the other hand, has only two basic elements in its lexicon, a high and a low. These can be used as pitch accents and edge tones, and combined into different tunes. But the variety of intonational patterns and meanings, while surprisingly rich, is still limited relative to words.

Gesture also has nearly unlimited expressive ability. Gestures can use three dimensions through time to express thoughts, and are not limited to certain words invoked in a linear order. Furthermore, gesture can use multiple articulators simultaneously: two hands, a head, and a face, for starters. Intonation, however, is limited to one articulator—the larynx—which can produce a signal in only one dimension—higher or lower glottal frequency—through time.

Yet despite the relative impoverishment of intonational meaning, I did find at least seven different pragmatic functions performed by co-occurring intonation and gesture. I'll describe each in turn, with examples.

Completeness. One of the most well-known pragmatic functions of intonation is signalling completeness with low phrase-final edge tones. Gesture can signal completeness in a number of ways: a simple one is to merely drop the hands to a resting position. In the "cupboards" clip, the subject explains her troubles fitting cupboard doors with the phrase "*The hinges are different.*", ending in a L-L% (low phrase accent and low boundary tone). Simultaneously, after making an iconic gesture for hinges, she drops her hands to her lap, as shown in Figure 15.

Figure 15. Example of gesture signalling completion, in conjunction with L-L% tones.

Incompleteness. Similarly, intonation can signal incompleteness, with a high edge tone. The "sous-chef" subject uses an H- phrase tone on the parenthetical "*who's a chef*" within the

longer series of utterances "*This guy is, he's a sous chef, uh, Jeff's sister's married this guy RJ, who's a chef, and is supposedly opening up this restaurant in DC, and this guy is his sous chef*". The tone is plateau-like, and the sense is that the speaker has more to say immediately afterwards, as if he is trying to get past the parenthetical, or is reporting a 3rd-party fact he isn't sure of, before returning to his main topic. The accompanying gesture reflects this incompleteness. For each person mentioned in this part of the discourse, the speaker makes a distinctive abstract deictic. In the left-hand picture in Figure 16, the speaker is pointing on the word *sister*. Later, in the right-hand picture, the speaker is pointing to a different location on the second mention of *this guy*. The middle picture shows the gesture on the word *chef*, accompanying the intonational plateau. Instead of a definite pointing, the speaker opens his hand and waves it slightly to his left. The sense is partly that of an emblematic shrug, and partly that he is offering up something temporarily and incompletely, in contrast to the surrounding definite pointing gestures.

Figure 16. Example of gesture signalling incompletion, in conjunction with an H- tone.

Figure 16 also illustrates two more cases of pragmatic synchrony between gesture and intonation: information status and focus.

Information status. Information status refers to the status of a discourse entity relative to the discourse. There is a rich variety of information statuses, and an extensive literature on the topic. Most researchers, however, agree that one status is a sense that an entity is somehow "new" to the discourse. As discussed in Chapters 2 and 3, a high pitch accent (H*) can signal that the accented word is new to the discourse. In the above "sous-chef" example, the speaker puts a high pitch accent on the word *sister*, which fits in that this is the first mention of the sister. As can be seen in the left-hand picture in Figure 16, the speaker also introduces this entity gesturally, by placing it in space with an abstract deictic.

Focus. In the same "sous-chef" example, the speaker has in fact nested parentheticals. He mentions *this guy*, then digresses to say how *this guy* is is related to someone known to the listeners. The digression includes a person named RJ. The speaker then digresses within the digression to explain who RJ is. After all this nested digression, he brings *this guy* back to the foreground ("*and **this guy** is his sous chef*"). The second mention of *this guy* receives focus both intonationally and gesturally. Intonational can signal focus with a high pitch accent, and the speaker places one on *this guy*. Gesturally, the speaker signals focus with not just a pointing gesture, but a more emphatic pointing gesture which drops his hand lower than the earlier deictics.

Related to focus are two more pragmatic functions which I found served by gesture and intonation simultaneously. These are emphasis and contrast.

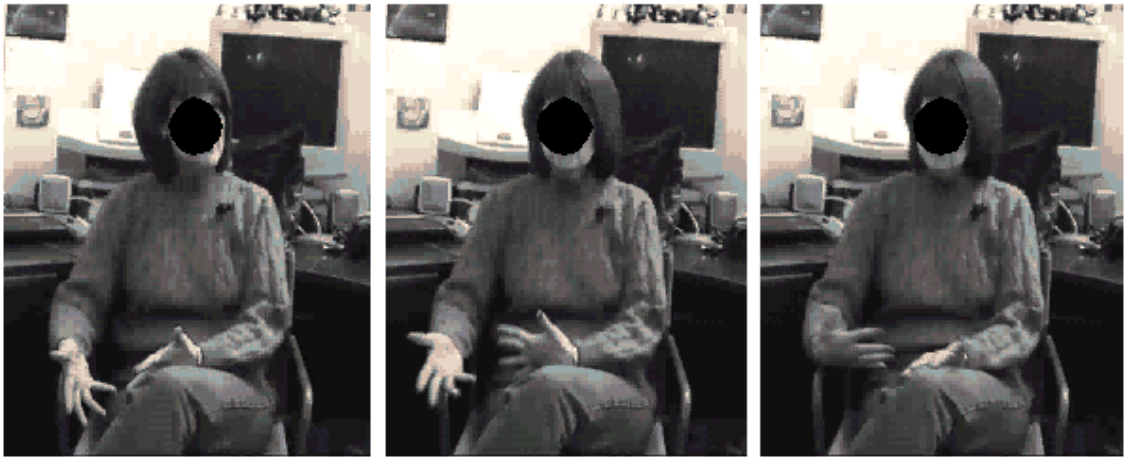
Emphasis. Whereas focus means to draw attention to a specific entity, emphasis is more general, and simply means that the emphasized entity is important. Although a high pitch accent can be used for emphasis, even more distinctive is an L+H*, which has a steeper rise to the peak. The "musicians" subject used an L+H* on the second mention of *know* in "*... know you're right <0.3 second pause>* ***know*** *that you're right ...* ". At the same time, the speaker brings his hands down forcefully on the word *know* (while retaining the handshapes he was using to indicate preciseness of musical tuning). This can be seen in the sequence of images in Figure 17. Although the downstroke looks slight in the still images (I've drawn a baseline for reference), it's quite clear in the video, as it takes place rapidly (within eight frames).

Figure 17. Example of gesture signalling emphasis, in conjunction with an L+H* tone.

Contrast. Contrast is also similar to focus, but whereas focus draws attention to one entity among several, contrast highlights a difference between the present entity or state and an earlier one. An L+H* pitch accent can also be used to indicate contrast, and was used by the "drywall" subject. She first discusses at length a remodeling in which walls, studs, and wires had

to come down. Her gesture during this discussion places "things coming down" to her right. This can be seen in the first picture in Figure 18. Subsequently, she continues: "*... and then, uh, he had to put the walls back* ***up*** *again ...*", with L+H* contrast on *up*. Her co-occurring gesture is shown in the second and third pictures in Figure 16. She first moves her hands even slightly more to the right, and then quickly shifts them to the left, to indicate the contrast of putting the walls back ***up***. Her hand orientations also reverse themselves, becoming almost the mirror image of what they were. Before the switch, her left hand was in a vertical plane and pointing to her right. After the switch, her right hand was in a vertical plane and pointing to her left. Because the gesture happens so quickly (within 4 frames), the sense is clearly one of contrast from a previous state.

Figure 18. Example of gesture signalling contrast, in conjunction with an L+H* tone.



Frame 754 (25:159)	Frame 825 (27:523)	Frame 829 (27:661)
"Things coming down", on speaker's right	Start of contrast, emphasize speaker's right	End of contrast, shift to speaker's left, reverse hand positions and orientations

Visual status of discourse. The final pragmatic function signalled by gesture and intonation is quite unusual and remarkable. The "sous-chef" subject gestured only with his left hand. His right hand stayed on the armrest, and occasionally made adaptor movements, rubbing the thumb and fingers together. Interestingly, this hand visually signalled the on-going status of the speaker's discourse. Before he started speaking, his hand was closed, as shown in the first picture in Figure 19. As soon as he started speaking, his hand opened up, as shown in the second picture, and stayed open precisely while he was speaking, as if to mean "the discourse is currently open". He did shift his index finger to the other side of the armrest while speaking, but his fingers remain open, as shown in the third picture. Interestingly, he made this shift during the parenthetical discussed above, as if the lesser discourse status of the parenthetical licensed the shift.

Figure 19. Example of gesture signalling discourse status, in conjunction with various tones.

Frame 0 (0:000)	Frame 24 (0:800)	Frame 277 (9:230)	Frame 308 (10:278)
Fingers together, discourse closed	Fingers apart, discourse open	After shift, fingers still apart, discourse still open	Fingers together, discourse closed again

As soon as he paused his speech, the speaker closed his fingers (shown in the fourth picture). Upon every subsequent resumption of speech, the fingers opened again, resembling the third picture, and then closed again upon every cessation of speech, resembling the fourth picture. Finally, part of the conversation consisted of a three-way interchange between the speaker and his

two listeners, with very short contributions by each. During this interchange, the speaker's fingers opened and closed repeatedly and rhythmically, but not quite in time with the conversational rhythm. The conversational rhythm during this interchange had a period of .8 seconds. The speaker's fingers made two open/close cycles with a one second period, and then four cycles with a period of .6 seconds. (The next section discusses rhythm in more detail).

Thus, the speaker's non-gesturing hand reflected the on-going status of his discourse: open, closed, parenthetical, and back-and-forth interchange. Intonation signalled the discourse status much less subtly. When an intonational phrase was under way, the speaker was obviously speaking. When it was finished, the speaker was pausing. The parenthetical was signalled by a reduced pitch range. Both modalities therefore indicated the meta-communicative status of the conversation.

To summarize this sub-section, gesture and intonation make occasional but definite joint contributions to pragmatic meaning. McNeill's pragmatic synchrony rule extends to intonation, with the caveat that the synchrony is intermittent, and not whenever the two modalities occur together.

5.6 How Do Gestural and Intonational Rhythm Correlate?

The final correlation I investigated between gesture and intonation is rhythm. Before I begin this discussion, let me define a few things.

First, the word *beat* can be overloaded when discussing gesture and rhythm. *Beat* can refer to a gestural beat; that is a certain type of gesture. *Beat* can also refer to a rhythmic beat (or metrical beat, or musical beat). Tuite (1993) called this a *rhythmic pulse*. To avoid confusion, in this section I'll use the full term *gestural beat* for the former, and Tuite's term for the latter.

Second, this discussion will rely on quite appropriate musical terms. Segments of a conversation will have one or more *tempos*[17]. There are various *instruments* (head, hands, larynx), which play *notes* (e.g. head nods, apexes, and pitch accents). The notes can be on the rhythmic pulse, or *downbeat*. Or, they can fall on the *backbeat* (the usually weaker upbeat between two stronger downbeats). Notes can also be *syncopated* (altered from the rhythm in a variety of ways). Notes can be doubled or halved, with twice or half the period, respectively. Finally, Anvil's musical score layout, which was adapted for speech and gesture research, will in this discussion be a true musical score.

Third, as mentioned above, there are a number of ways humans can play a "note". Humans can use head nods, or gestural strokes, or gestural beats, or pitch accents, or stressed syllables, or eyeblinks, or a variety of other phenomena. It can be cumbersome to refer to each of these separately, when in terms of rhythm they are all in the same category: a distinctive note expressed by the human, which may or may not fall on the rhythmic pulse. Therefore, I'll coin a new term for all of these phenomena collectively. I'll refer to them as *pikes*. This term is in honor of Kenneth Pike, one of the first to point out a possible relationship between intonation and

[17] For musicians, tempo is a perceptual abstraction. They need not play a note on the tempo, but they know what it is. The tempos I'll be discussing are empirical, derived from the data.

gesture. The word *pike* is also reminiscent of the words *peak* and *spike*, which roughly capture the feeling of these short, distinctive expressions. A pike, therefore, is a short distinctive expression, either in speech (a pitch accent or a stressed syllable), or in movement (a gestural beat, the apex of a stroke, a head nod, an eyeblink, or any other distinctive point). In the musical analogy I've used, in which the body and speech articulators are different instruments, pikes are the notes played, regardless of the instrument.

The idea that the various phenomena which I call pikes are more similar than different has been proposed by others. Tuite (1993, p. 100) proposed that all gestures are like gestural beats, but that representational gestures have other layers superimposed on the gestural beats. Susan Duncan (personal communication, August 26, 2002) has extended the idea of a beat to include not only all types of gesture, but also speech. To paraphrase her, "Everything's a beat". Like Tuite, she feels that all gestures are overlaid on gestural beats, and that "pure" gestural beats are merely gestures with nothing overlaid on them. In addition, stressed syllables and pitch accents are spoken versions of gestural beats. My term *pike*, therefore, may be a convenient term for what Duncan means by saying, "Everything's a beat".

Having presented some definitions, I'll now discuss gestural and intonational rhythm. My discussion will be in three subsections. First, I'll show examples from my data in which the two modalities both converge and diverge rhythmically, much like a jazz piece. Second, I'll statistically present the tempos used by the various instruments throughout my data. Finally, I'll discuss eyeblinks.

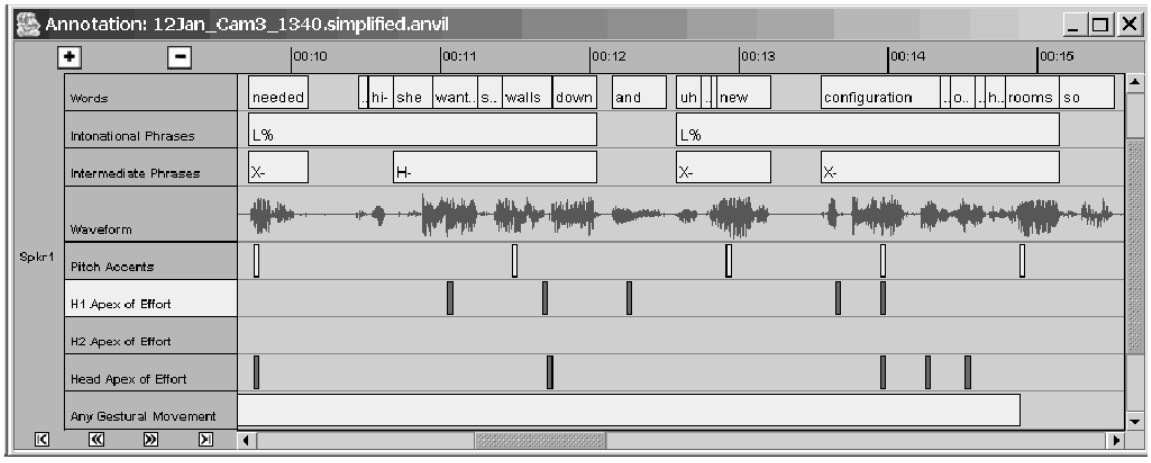
5.6.1 Jazz Music

Using Anvil's interface, I was able to look at the rhythmic score of the notes played by the three main instruments of head, hand, and intonation. Figure 20 shows an example from the "drywall" clip.

Before I discuss the rhythmic score in Figure 20, let me first explain what's shown in the figure. This is Anvil's annotation board, as seen before, but with any tracks unrelated to rhythm removed, for simplicity of exposition. Furthermore, Anvil has been zoomed out, to display more time in a given screenwidth. The top track in Figure 20 contains the words. Because Anvil has been zoomed out, they are not entirely readable. The speaker is saying, "*... needed <pause> uh, hi- she wanted some walls down, and, uh, like a new <pause> configuration of one of her rooms, so ...*".

The second and third tracks present the spans of intonational and intermediate phrases, respectively. Pitch accents will only occur during speech, so these phrases are included for reference, so the observer can deduce whether the lack of a pitch accent at a certain point is due simply to the lack of any speech at all. The next track displays the waveform, whose amplitude bursts give clues to spoken events. The next four tracks display the pikes I annotated: pitch accents, hand 1 apexes, hand 2 apexes, and head apexes. As discussed earlier, these are all points in time, but have been represented in Anvil with 1-frame spans, the right-hand edge of which is the actual point of the pike. Finally, the bottom track displays spans of any gestural movement. This helps the observer to deduce whether the lack of bodily pikes at any point is simply due to the lack of any movement at all.

Figure 20. Example of rhythmic score from the "drywall" clip.



In describing the rhythmic score, I'll start first at the bottom, with the head. This instrument starts out with a single pike, or note, which aligns exactly with a single pitch accent, on the word *needed*. The head has another single pike several seconds later, which aligns with a hand 1 apex, at the words *walls down*. The head ends this section with three notes about .3 seconds apart, the first of which aligns with both hand 1 and with a pitch accent, at the word *configuration*.

The hand (hand 1) starts out somewhat later than the head, with three pikes about .6 seconds apart, the middle one aligning with a head pike. It ends with two notes .3 seconds apart, the final one aligning with both the head and a pitch accent.

We can already see a rhythm between the hand and the head. The hand has an initial rhythm of .6 seconds, then doubles to .3 seconds, to be in rhythm with the head, although the hand stops "playing" when the head starts "playing" its final notes. Furthermore, the hand and head align at two points.

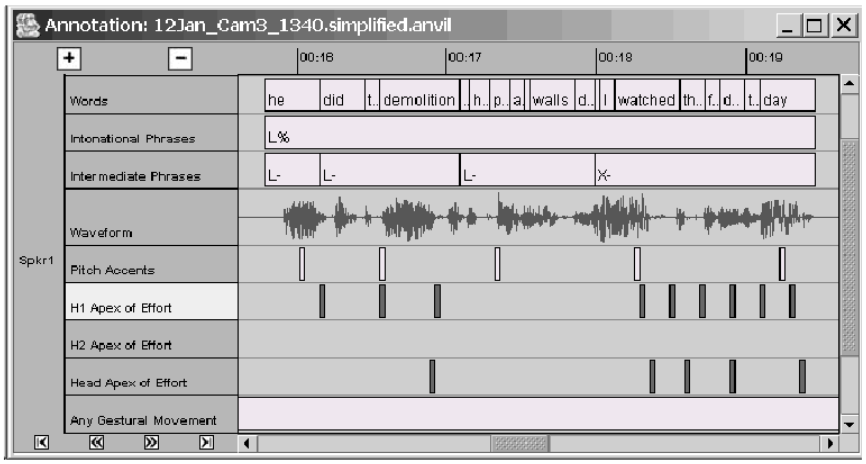
Now for the final instrument, the larynx, playing pitch accent notes. Its first pike aligns with the head. The second pike aligns with neither the head nor the hand, though it could be a

syncopated note, slightly anticipating the joint note of the head and hand. The final three notes have a tempo of 1 per second, which is not an exact multiple of the head or hands (which play notes every .3 and .6 seconds). But the middle of the larynx's final three pikes aligns with the other two instruments.

Thus, we have what could be a score for jazz music. The three instruments sometimes play their own individual tempo, and sometimes play a common one. They sometimes align on certain notes, and usually don't. They may even syncopate from the rhythm. They are clearly inter-related, though not uniformly so.

As another example, Figure 21 shows the "drywall" section immediately following the above example.

Figure 21. Second example of rhythmic score from the "drywall" clip.



In Figure 21, the words are "*he did the demolition, so he pull* [sic] *all the walls down, and I watched this from day to day*". In this section, the jazz score is a bit more complicated. The head starts with a single pike, which aligns with the hand. Then it has four final notes, which appear to be slowing (the intervals are .23, .30, and .47 seconds, respectively), perhaps to match

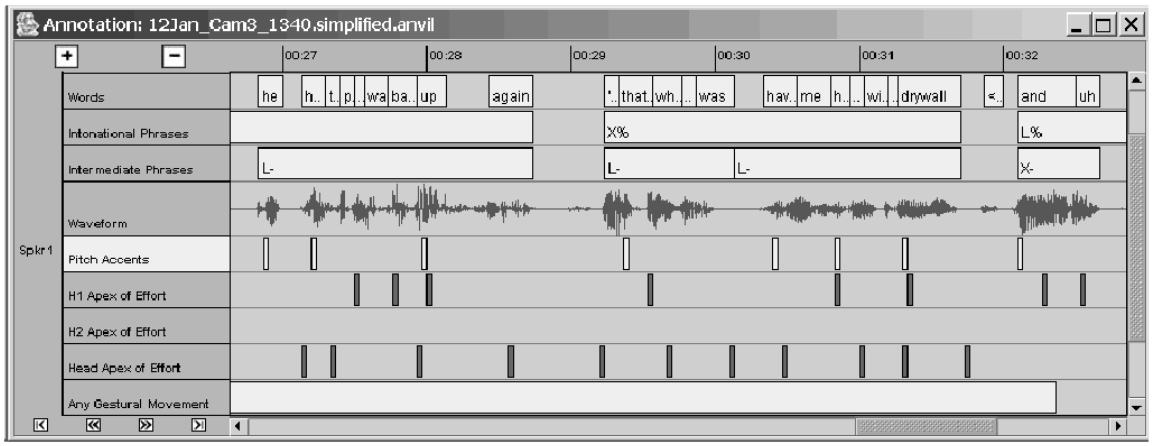
the slowing of speech on the phrase-final word. The second-to-last head pike aligns with a hand note, and the last note is probably meant to align with the final hand and pitch accent notes.

The hand, as in the previous example, doubles its frequency, starting with three pikes .4 seconds apart, then six pikes .2 seconds apart. These final notes do not slow at the end.

The larynx also appears to be slowing throughout the piece, with inter-pike intervals of .54, .77, .93, and .97. The pitch accents align with other pikes in several places.

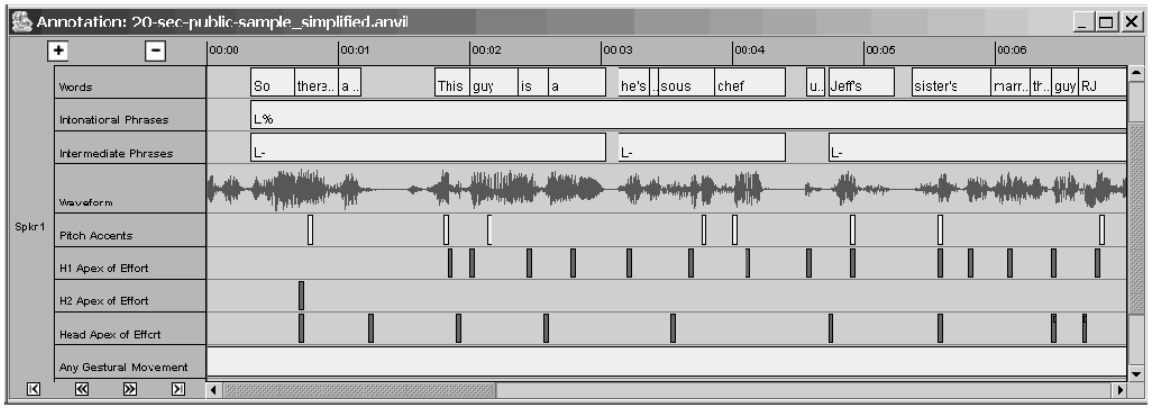
Figure 22 shows a third example from the "drywall" clip, included to show that the rhythmic relationships are not always as clear.

Figure 22. Third example of rhythmic score from the "drywall" clip.



While there is some regularity within each separate track in Figure 22, there is also a great deal of irregularity. The unifying feature is that there are several points where all three modalities converge. Indeed, this is what typically seems to occur in my data: two or three "strong" points where all three instruments join in, often at the start and end of sections of discourse. Figure 23 gives a clear example of this.

Figure 23. Example of rhythmic score from the "sous-chef" clip, showing re-setting of rhythm.

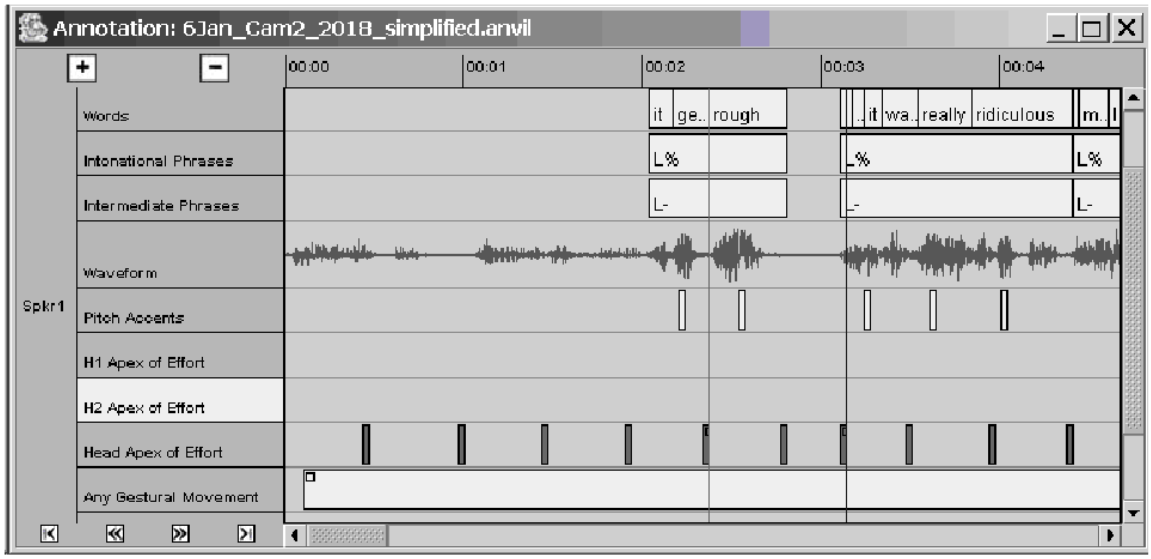


In Figure 23, the subject starts his story by introducing the sous-chef ("*So there's a …*"). The speaker sets a strong rhythmic point to start with, with three of the four instruments contributing a pike on the word *there's*. Then the speaker is disfluent, as he makes several false starts digressing to explain how the sous-chef is related ("*This guy is a … he's a sous-chef … uh, Jeff's sister's married this guy RJ …*"). During the disfluency, the three main instruments are very arhythmic. By the time he reaches the word *sister*, the speaker has regained his stride, and sets a strong rhythmic point on *sister*, again with three pikes, to get his rhythm going again.

A speaker's rhythm can be extremely regular. Figure 24 shows the first five seconds from the "musicians" clip. Note the head pikes. The speaker is initially nodding in assent to his listener's comment that it can be difficult to work with fellow musicians. Then, at the point of the first vertical bar, he switches from head nods to head shakes, to accompany "*it gets rough*". Then, at the point of the second vertical bar, he switches back to head nods to accompany "*I mean, it was really ridiculous*". The head pikes are almost perfectly isochronic throughout—just under a half-second per interval—except for the transition from the shake to the second nod, where there are two intervals closer to a third of a second. The pitch accents at first alternate

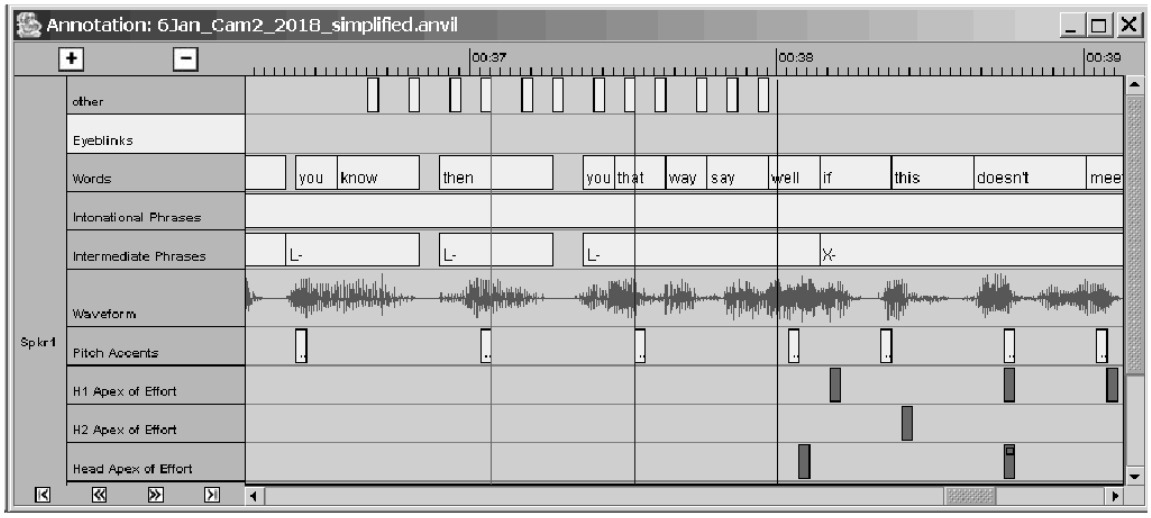
neatly with the head pikes, then the two converge (within several frames) on the stress in *ridiculous*.

Figure 24. Example of rhythmic score from the "musicians" clip.



The same subject later had a speech disfluency, during which he started moving his leg in a sewing-machine fashion (foot on the floor, knee at a right angle and moving up and down rapidly). As the leg movements were in time with other pikes, I annotated them just for this section. This sequence is shown in Figure 25.

Figure 25. Second example of rhythmic score from the "musicians" clip, showing leg synchrony.

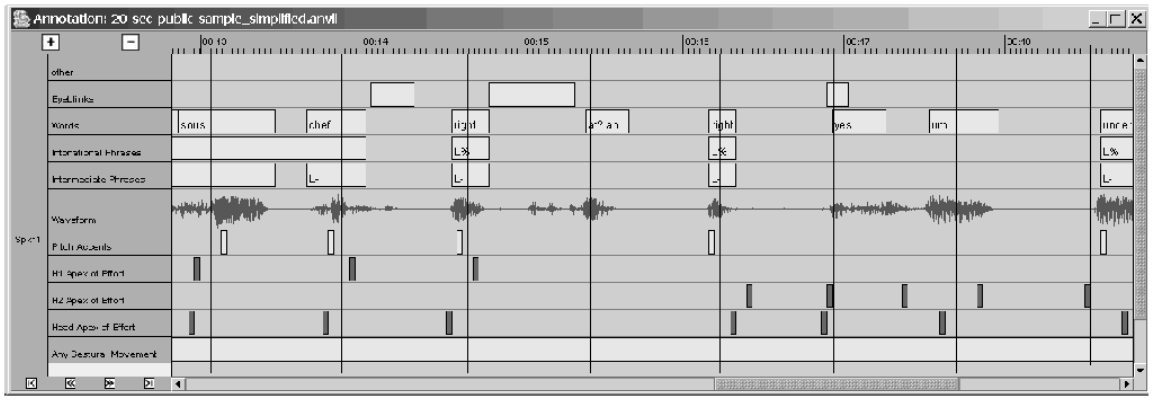


The top track, labelled "other", contains the leg movements. Every two leg pikes is a full up-down movement cycle. The subject makes six full cycles, each lasting six or seven frames. As can be seen by the vertical lines, the downstroke of every other cycle coincides with a pitch accent. The last downstroke also coincides with a head pike. Although the last leg pike is slightly off from the pitch accent and head pike, it is still within four frames, and is perceived as nearly simultaneous in the video. The leg motion stops when the disfluency ends, on the word *well*. As in previous examples, the subject resets his rhythm with a strong three-pike rhythm point on *doesn't*.

As a final example of rhythm, I'll discuss an interesting case of inter-speaker rhythm. I've already mentioned that the "sous-chef" clip was selected because of interactional synchrony between the speakers, even though interactional synchrony was not a focus of my research. When playing the three-subject video slowly back and forth, one can see the three participants slightly raise and lower their heads in synchrony, in a way which is difficult to reproduce with still images. However, there was another interesting point about this clip, which I can display.

That is the inter-speaker rhythm during a series of short interchanges between the three, as shown in Figure 26.

Figure 26. Example of inter-speaker rhythm, from the "sous-chef" clip.



In this interchange, the speaker says the following eight words, at nearly regular intervals of approximately three-quarters of a second: "*sous ... chef ... right ... at?/and? ... right ... yes ... um ... under*". These stressed syllables can be seen in the amplitude bursts on the waveform. I've drawn vertical lines at what appear to be the correlation of these speech pikes with other pikes. Interestingly, the speaker is interrupted three times by his listeners. During the speaker's word *chef*, one listener simultaneously overlaps with the words *sous-chef*, which causes the speaker to reply *right*. During the unintelligible word *at* or *and*, the same listener simultaneously overlaps with the words *number-two chef*, which causes the speaker to reply *right* again, followed by *yes*. During the speaker's subsequent *um*, the second listener simultaneously overlaps with the words *is that what sous means*, to which the speaker replies *under* (as in, "*sous* means *under*"). The interesting point is that all these interruptions occur exactly on the speaker's rhythmic pulse. The listeners interject their contributions of two, four, and five syllables quite quickly, matching

the speaker's slowly spoken one-syllable contributions. It's as if the listeners want to jump in and out of the conversation without upsetting the speaker's rhythm.

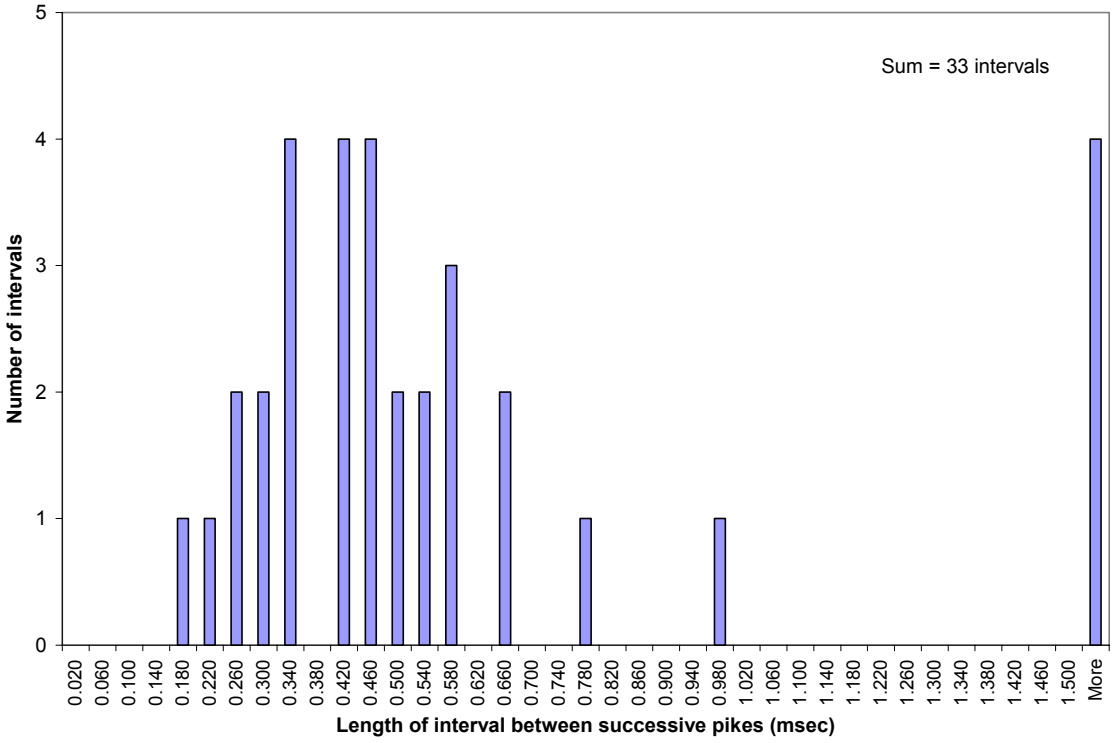
To summarize this sub-section, the different articulators, or instruments, of body movement and intonation are related rhythmically. The relationship is not uniform, only manifesting at certain points, but is nonetheless quite clear.

5.6.2 Tempos

The examples in the previous sub-section showed varying intervals between pikes, depending on the speaker, the articulator, and differing sections in the data. Are there any common tempos for certain articulators? For certain speakers?

As the annotations are all time-stamped, it's relatively easy to answer these questions. For each articulator, and for each speaker, I captured the interval between successive pikes. I then plotted a histogram of these intervals, grouping them into bins of 40 msec. Figure 27 shows one such histogram, for the "sous-chef" speaker's hands.

Figure 27. Intervals between hand pikes, from the "sous-chef" clip.



As can be seen, most of this speaker's intervals between successive hand pikes were in a range between .18 seconds and .66 seconds, approximately. That is, his hands typically had a rhythm with intervals (or periods) somewhere between .18 and .66 seconds. This would mean a tempo (or frequency) range between approximately 5.5 pikes per second and 1.5 pikes per second. (The tempo (or frequency) is the inverse of the interval (or period)). This range could accommodate doubling of notes, as we saw in some of the above examples. That is, the speaker could double notes from an interval of .6 down to .3, or from an interval of .4 down to .2.

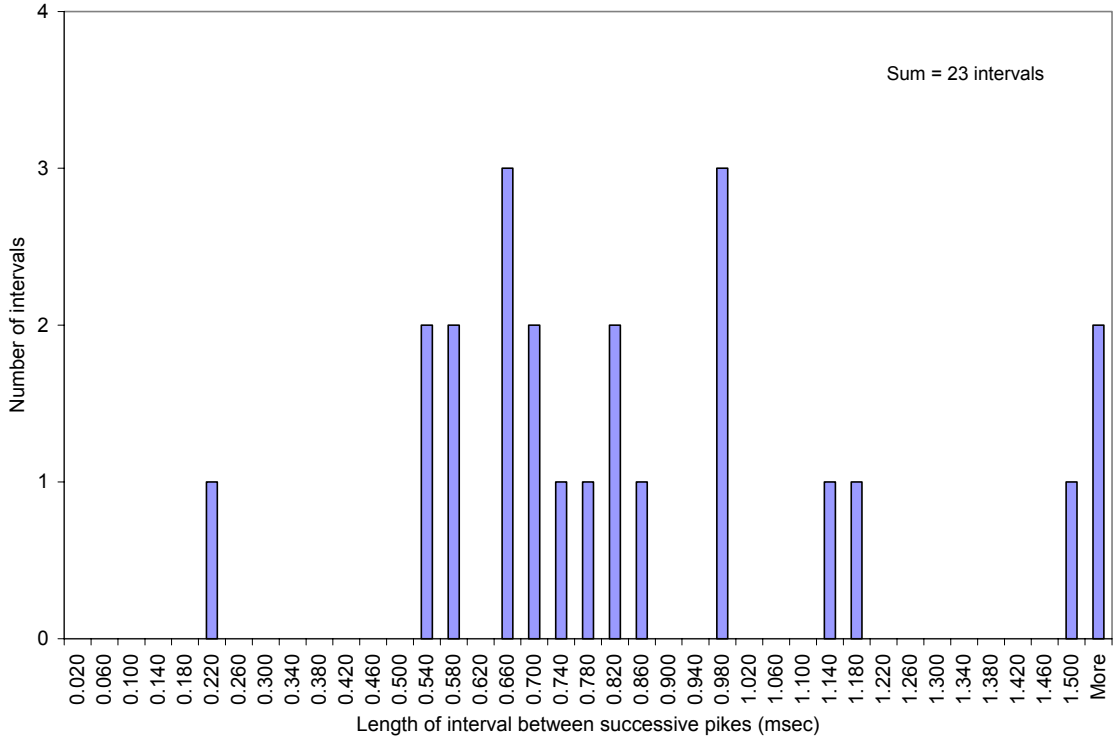
Several explanations are in order regarding this and subsequent histograms. There are only 33 intervals measured in Figure 27, so care should be taken in inferring too much; the figure can only be used for general trends.

Note also that there are four intervals of more than 1.5 seconds, on the right-hand side of the histogram. These are intervals covering longer stretches of no hand activity. I originally tried to only measure tempos during periods of clear activity of the hand (or head, or larynx), thus ignoring the intervals between such periods. However, it became extremely tricky to decide by eye just which intervals belonged to a period of activity within a cohesive group of pikes, and which intervals were intervals between those periods of activity. Therefore, I let the data stand as is, and simply plotted all intervals, regardless of length.

One more explanation is in order. The histogram in Figure 27, and all subsquent hand-interval histograms, contain data from both the subject's hands. However, the intervals are only measured between successive pikes on the same hand. I combined the data from both hands in the histograms, to try and get a more complete picture the tempo of the hand in general.

Having presented a histogram of hand intervals, we can now compare it with a histogram of head intervals, from the same subject (Figure 28).

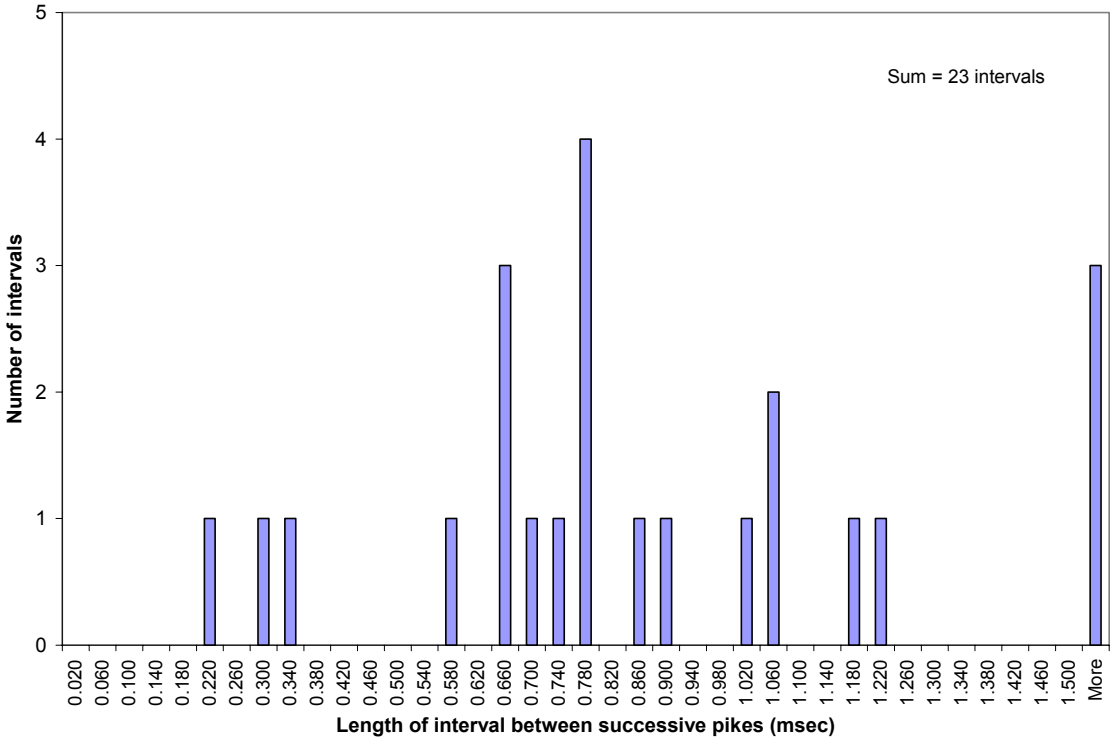
Figure 28. Intervals between head pikes, from the "sous-chef" clip.



This speaker's head intervals were in a higher range than his hands. His head pikes were spaced in a range approximately from .54 seconds to 1.18 seconds, with a possible cluster around 2/3 of a second. This is different than the hand histogram, where intervals were faster, clustering possibly around .4 seconds.

While the head looked different from the hands in terms of typical inter-pike intervals, the head looked more similar to speech. Figure 29 shows the speech interval histogram for the same subject.

Figure 29. Intervals between speech pikes, from the "sous-chef" clip.



As can be seen, the range of intervals for speech is more similar to that of the head (clustering well above a half-second) than it is to that of speech (clustering below a half-second). The clustering here is admittedly less clear than in the histograms for hand and head pikes, with fewer histogram bins containing more than one interval.

I should explain that the speech intervals shown in Figure 29, and all subsequent speech-interval histograms, are intervals between successive stressed syllables, and not strictly pitch accents. The relationship between pitch accents and stressed syllables is of course a close one: pitch accents go on stressed syllables, but occasionally stressed syllables will go without a pitch accent, if there is no prominent pitch movement nearby. In my data, there were relatively few stressed syllables without pitch accents, so the histogram largely represents that of pitch accents by themselves. I included stressed syllables in my histograms for two reason. First, other studies

of communicative rhythm have used stressed syllables, and I wanted to be consistent with the literature. Second, it seemed clear in several cases that the rhythm of speech could be better described by including stressed syllables as well as pitch accents. A case in point is the example I described in Figure 26 above, where the listeners took care to interrupt on the rhythmic pulse. Of the eight isochronic speech bursts made by the speaker, there were pitch accents on only five of them. I felt that recording stressed syllables would more completely capture the speaker's rhythm of speech.

I've so far shown the typical rhythmic intervals for the hands, head, and speech of one speaker. How do the intervals for the other speakers look? And for all the speakers combined?

Figure 30 displays all such histograms together. Looking down the columns, one can observe how the various articulators' rhythms line up, for each individual speaker and for all speakers combined. Looking across the rows, one can observe how the same articulator varies across speakers. Due to the small size of each histogram, I have omitted the axis labels; they are the same as in Figures 27-29 above. To aid the reader, I have drawn vertical lines through each histogram at 0.5 and 1.0 seconds.

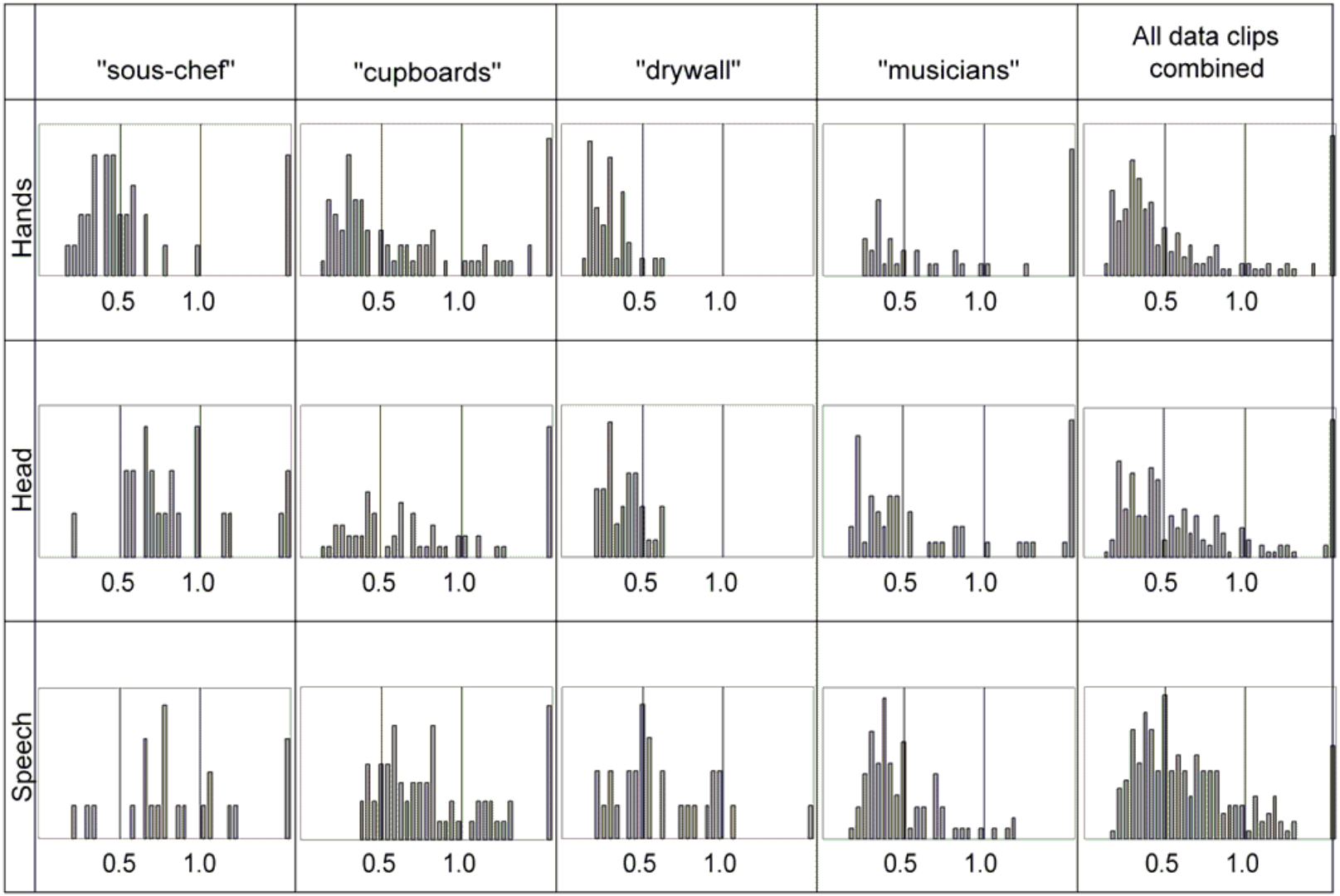
As can be seen in Figure 30, although there is some variability, for most individual speakers and for all speakers combined, the intervals for speech (bottom row) are longer than for the hands (top row). The intervals for the head (middle row) are somewhere between the other two. Perhaps the hands produce pikes at shorter intervals because they're more nimble and expressive, while the less nimble head requires longer intervals between pikes. Speech pike intervals may be longer still due to pulmonary constraints.

The "musicians" subject is an exception: all three articulators are closely aligned. And in terms of range, the "drywall" and "musicians" subjects have smaller intervals, a half-second or

less for all modalities. The “sous-chef” and “cupboards” subjects have larger intervals for the head and speech, a half-second or more.

Regardless of the individual differences, the main point of Figure 30 seems to be that, whatever the cause, each articulator appears to have a slightly different rhythm.

Figure 30. Intervals between pikes, for all articulators and all speakers. See Figures 27-29 for axis labels.

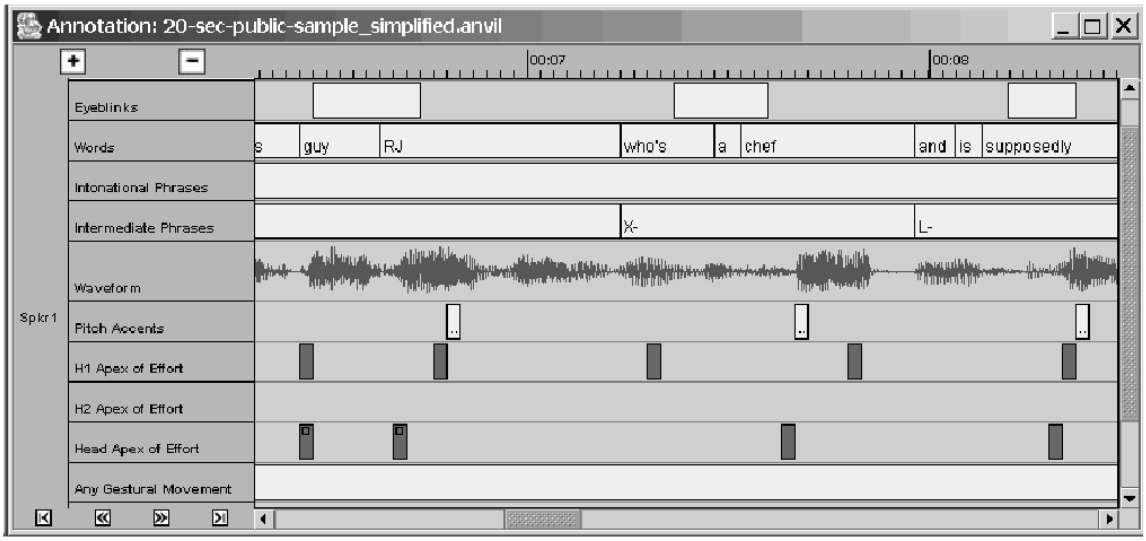


5.6.3 Eyeblinks

In investigating rhythm, I noticed that eyeblinks appeared to be occurring in time with speech and the rest of the body. This is not a new discovery. As mentioned earlier, Scheflen and Birdwhistell observed that eyeblinks occur at intonational junctures (the end of a clause) along with head nods and hand movements (Scheflen 1964, pp. 320-321). And Condon and Ogston (1967, p. 229) noted that "the eye blink … also functions as an accentual parakinesic phenomenon.… eye blinks do not seem to occur randomly, but are also related to other ongoing variations in the sense that *if* they occur, their point of occurrence may be relatively specifiable" (italics in original). I decided to annotate the eyeblinks in my data, and see how they fit in with the other articulators.

I found that eyeblinks typically happen on the rhythmic pulse. A casual viewing of my video data (or of anyone speaking) will confirm this. Eyeblinks co-occur not only with stressed syllables, but with bodily pikes as well. Even more intriguingly, upon close examination, eyeblinks don't typically happen *on* the rhythmic pulse, but just prior to it, so that the eyelids are re-opening on the rhythmic pulse. It's as if eyeblinks are a syncopated note, slightly anticipating the rhythmic pulse. Figure 31 shows an example of this, from the "sous-chef" clip.

Figure 31. Eyeblinks from the "sous-chef" clip, with eyelids re-opening on the rhythmic pulse.



In Figure 31, the top track shows eyeblinks. I annotated each blink from the start of eyelid closure until the end of re-opening. While I took other pikes to be a definite instant in time, I wasn't sure a priori which point of the eyeblink to call the most important point. The moment of closure seemed a logical choice, but it also seemed to me that the re-opening was timed to occur with other pikes. Therefore, I simply coded the entire interval of the eyeblink, as shown in Figure 31.

As can be seen, each eyeblink is timed so that it ends (i.e. the eyes are re-opened) with other pikes (including a waveform burst), on a rhythmic pulse. This is very common in my data, and three out of four subjects timed most of their eyeblinks thus. It's almost as if the speaker were holding the eyes closed until the rhythmic moment, and then opening them, just as manual gestures hold their position, and then perform the stroke at the appropriate moment. In terms of manual gestures, then, the closing of the eyelids would be the preparation, the period of closure would be the hold, and the re-opening would be the stroke. It's interesting that most eyeblinks in my data took longer than the minimum apparently needed to moisten the eye. The minimum

eyeblink in my data lasted three frames (100 msec), yet the average was six (200 msec). The extra time could be used for the hold, to wait for the appropriate moment to re-open.

The fourth subject, the "musicians" speaker, was an exception. His eyeblinks were timed to straddle the rhythmic pulses, rather than anticipate them. But the eyeblinks occurred on the rhythmic pulse nonetheless.

Like the other pikes, eyeblinks showed a remarkable adherence to rhythm. Figure 32 shows a clip from the "cupboards" data.

Figure 32. Eyeblinks from the "cupboards" clip, showing adherence to rhythm.

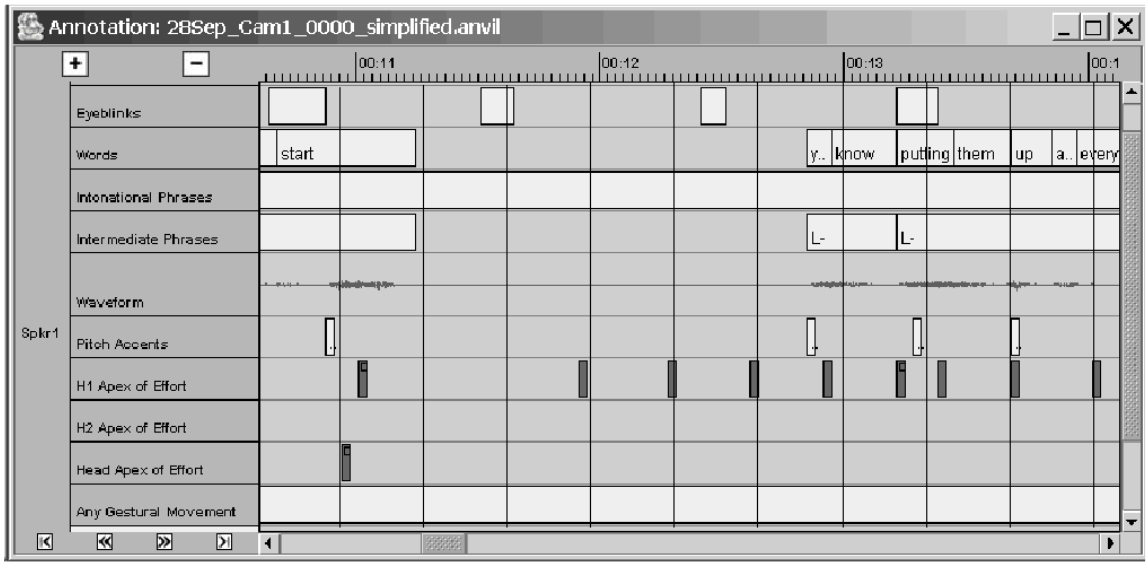


Figure 32 shows a part of the speaker's utterance "*and so I start <1.6 second pause> you know, putting them up and everything*". As discussed earlier, during the pause, the speaker silently places four imaginary cupboard doors in the air. I've drawn isochronic lines through Figure 32, roughly 10 frames apart. As can be seen, all events but one happen near a rhythmic downbeat. Before the pause, the speaker sets a strong point with three pikes and an eyeblink on the word *start*. Exactly two rhythmic pulses later, she blinks. The following three rhythmic

pulses are accompanied by hand pikes. The final four rhythmic pulses mark time to the words "***you*** *know,* ***put****ting them* ***up*** *and* ***ev****erything*", with a hand pike and/or a pitch accent close by. There is also an eyeblink on ***put***. When viewing this clip in slow motion and counting the meter out loud, one can see the speaker move in rhythm.

The only event not on a rhythmic downbeat is the third eyeblink. But this happens precisely on the upbeat between two downbeats. The reason, I believe, is that during this silent section the speaker is actively gazing at her hands as they position the cupboard doors. To avoid having her eyes closed for the main communicative event during this silent period, she blinks on the upbeat, between hand pikes. This blink, three frames long, is also the shortest in my data.

Later, the same subject, while listening to her conversational partner, blinked in rhythm with her *partner's* speech. The partner was asking "*and then you have to* ***lift*** *it up, and like* ***hold*** *it?*", with stress on *lift* and *hold*. The original subject (who was now listening) blinked exactly on these two words.

I initially planned on investigating the rhythmic relationship between intonation and gesture, looking at pitch accents and manual gestures. I then extended the study to speech in general, including stressed syllables, and to more of the body, including the head and even the leg. And now it seems even eyeblinks obey the body's rhythm. The rhythmic relationship between intonation and gesture may be a manifestation of a much deeper phenomenon.

5.7 Comparison of Findings with Others in the Literature

Micro-analytic studies of gesture and intonation are still uncommon. Therefore, it would be interesting to see how my findings compare with others in the literature. This will also be a convenient place to specifically answer Yerian's questions, which inspired my research, and I'll begin with these.

I'll reprint Yerian's questions (Yerian 1995, pp. 15-18) as posed in my introduction chapter, by paraphrasing them one at a time in italics, and then answering them.

The proposal that gestural movement may at times parallel the direction of pitch movement may be looked at more closely beyond terminal rises and falls. When there is parallel movement, does it coincide with certain pitch accents or phrases? As discussed, I found no evidence for parallel movement of gesture and pitch.

Are there patterns of body movement which correspond to pitch accents? Yes, apexes of gestural strokes correlate clearly with pitch accents.

Are there patterns of body movement which correspond to intermediate phrases? Yes, gestural phrases typically, but not always, correlate with intermediate phrases.

Are there patterns of body movement which correspond to intonational phrases? No, not directly. But gestural phrases do correspond indirectly to intonational phrases, in that gestural phrases correspond with intermediate phrases, which make up intonational phrases.

Are the holds of gestures sensitive to any or all of these boundaries? Occasionally, gestural phrases align with intermediate phrases only if the post-stroke hold is disregarded. In these cases, the hold is sensitive to an intermediate phrase boundary.

When more than one gesture occurs in one intonational phrase, do these gestures show sensitivity to intermediate phrase accents? Yes, this is another way of saying that gestural

phrases correspond to intermediate phrases. So more than one gesture occurs in an intonational phrase, with sensitivity to intermediate phrase accents.

Are holds dropped before, during, or after boundary tones? By "holds dropping" near a boundary tone, I assume Yerian means holds transitioning to a retraction at the end of an utterance. There was no clear pattern in my data as to the timing of pre-retraction holds with nearby boundary tones.

Might Pierrehumbert's hypothesis that intermediate phrases are characterized by 'catethesis' find support in some parallel form of gestural movement? No, I found no parallel gestural form of catathesis (or downstep) in concert with intermediate phrases, or at all. It's difficult to know a priori what catathesis would look like gesturally, but in addition to looking for any pattern whatsoever, I also checked specifically for successive movements dropping lower down, or having reduced effort. I found no such pattern.

Gestures may align themselves differently in response to the various complex pitch accents Pierrehumbert proposes. I found no such correlation.

I'm happy to say that I was able to answer all of Yerian's questions, though the answer to many of them was "no". I'll now compare my findings with those of others who studied gesture and intonation, beginning with Kendon.

As discussed, Kendon (1972, 1980) proposed a relationship between gesture and intonation on five hierarchical levels. First, he noticed that strokes align with stressed syllables. This I confirmed, with the slight variation of looking at apexes of strokes and pitch accents.

Second, Kendon aligned gestural phrases with tone units. Assuming tone units would correspond to Pierrehumbert's intonational phrase, I found that gestural phrases often correspond to a smaller unit, the intermediate phrase. Yet I feel that Kendon's observation fits my data in spirit, for several reasons. Often there is only one intermediate phrase in an intonational phrase.

In addition, the distinction between an intermediate and intonational phrase is not always clear. Finally, Kendon's use of tone units preceded Pierrehumbert's idea of a smaller intermediate phrase, so Kendon used the most appropriate theoretical construct available to him.

Third, Kendon aligned gestural units with locutions, or sentences. I found only an approximate correlation here. Sometimes there was a one-to-one relationship, but often a gestural unit would span many syntactic sentences.

Fourth, Kendon noted that groups of gestural units sharing consistent head movement corresponded to locution groups, or locutions sharing a common phonological feature (e.g. all ending with a *low-rise* contour). Although I wasn't initially looking at units above the intonational unit, nor at common head movements over time, in going back to check Kendon's claim, I found it to be largely true. For example, the "cupboards" subject produces two successive intonational phrases, separated by a pause, each with the distinctive set of tones L* H-L%. These phrases are "... *because they're all white <pause> I think the lines really show*". On each phrase, she lowers her head down and to the left, while her eyes look up and to the right at her listener.

Kendon's fifth claim was that consistent arm use and body posture were synchronized with a locution cluster, or paragraph. As discussed, McNeill refined this to show that recurrent gesture forms, or catchments, were used within discourse segments. Again, though I wasn't initially investigating such a large hierarchical level, upon re-checking my data, I found McNeill's refinement to be true. As to body posture, none of my subjects made major posture shifts, so this claim couldn't be verified.

Overall, Kendon's five observations were largely borne out by my data, with the caveats discussed. I'll now turn to some of McClave's (1991, 1995) findings.

McClave specifically tested Bolinger's parallel hypothesis by looking at final tones and their correlation with movement, and failed to find a correlation. As discussed, I found no correlation either.

McClave found that gestures don't cross tone unit boundaries, indirectly supporting McNeill's observation that gestures don't cross clause boundaries. Since I found that gestural phrases typically pattern with intermediate phrases, which make up intonational phrases, I also confirmed this pattern.

McClave's beat hypothesis stated that groups of beats existed in which one beat was anchored on the intonational nucleus, while the other beats were isochronically spaced from the anchor, and not necessarily on stressed syllables. I also saw this pattern.

McClave noted gesture fronting, in which multiple gestures in a single tone unit were shortened and started earlier. This correlates to my finding that multiple gesture phrases could exist in a single intermediate phrase. In these cases, the gesture phrases were indeed shorter and started earlier.

Thus, like Kendon's findings, McClave's findings were also largely supported by my data. I'll now look at previous work on speech-gesture rhythm.

As discussed, Condon (1976) proposed a "rhythm hierarchy", with five levels: phone, syllable, word, half-second cycle, and full-second cycle. As to rhythm at the level of phones, I didn't code my data to such a fine-grained distinction, limiting my annotations to the word (and sometimes the syllable) level. I therefore can't comment on phone-level rhythm. As to the syllable-level rhythm, I certainly found this to be the case. However, I did not notice a word-level rhythm. I believe rather that it's the prominent syllables in certain words that pattern rhythmically.

As for the half-second and full-second cycles, these certainly fall within the range of tempos I measured (shown in Figure 30), but I didn't find these specific tempos existing apart from others. As discussed, McNeill (1992) measured successive strokes roughly one to two seconds apart. My tempos were a little quicker than this, although I was measuring apexes, not strokes per se, and I often had multiple apexes for multi-part strokes.

In terms of rhythm, then, I concur with Condon that rhythm exists, but I'm not able to pin down a tempo, or a hierarchy of tempos, as precisely as he has. Like Condon, I also noticed interactional rhythm, between conversational participants.

My final comparison is on eyeblinks. Condon and Ogston (1967, p. 229) noted: "The eye blink has been found to occur during vocalization at the beginning of words or utterances, usually with the initial vowel of the word; at word medial syllabic change points; and precisely following the termination of a word." In other words, they saw eyeblinks occurring on word boundaries—either at the start of a word, or at the end, or at a word-internal syllabic boundary. My observations don't exactly match Condon and Ogston's. As mentioned, I found eyeblinks to occur on or just before the rhythmic pulse, and found no correlation with word boundaries. If a word boundary happened to fall on the rhythmic pulse, then yes, our observations happen to match.

Another problem is that I coded the entire duration of the eyeblink. These often lasted several hundred milliseconds, as long as many words. This granularity is too coarse for Condon and Ogston's fine-grained claim, and I presume they were annotating the exact moment of closure.

6 Discussion

When presenting my results in the previous chapter, I said little about their theoretical implications. I'll discuss those implications in this chapter. I'll begin by briefly recapping my findings.

6.1 Recap of Findings

What can we say about the relationship between gesture and intonation? I can only make claims about my four subjects; I cannot generalize to any given population. Yet the findings seem to indicate that there is a definite relationship between the two modalities, which manifests in a variety of ways. Here are summarized the positive findings from my data, which describe that relationship.

- Gestural and intonational events, regardless of type, cluster near each other on average.

- Apexes of gestural movement are aligned with pitch accents.

- Gestural phrases (g-phrases) are aligned with intermediate phrases. Typically one but often several g-phrases will align with a single intermediate phrase. This means that g-phrases respect intermediate phrase boundaries. When more than one g-phrase occurs in an intermediate phrase, the g-phrase boundaries are at syntactic and/or pause breaks in the intermediate phrase. Occasionally, a g-phrase will align with an intermediate phrase only when the post-stroke phases are disregarded. G-phrase boundaries typically slightly precede their corresponding intermediate phrase boundaries.

- Gesture and intonation can express the same pragmatic meaning in occasional but definite ways.

- Gesture and intonation are rhythmically related. While they each typically have their own tempo, the tempos are often interrelated, and they often join on rhythmic pulses. The range of tempos differs for the various articulators. The hands typically mark pikes at intervals less than 0.5 seconds, speech typically does so at intervals greater than 0.5 seconds, and the head's intervals are typically somewhere in between. The rhythmic relationship extends to eyeblinks, and to interactions between conversational participants.

In addition, I'll recap two negative findings: i.e. theories I specifically investigated but for which I found no evidence.

- Parallel vertical movement of gesture and pitch was not supported.

- No correlation between gesture type and tone type was found.

In short, gesture patterns with intonation in timing, structure and meaning.

I'll now turn to the theoretical implications of these findings.

6.2 Theoretical Implications

In terms of gesture theory, the relationship between gesture and intonation lends support to the theory of a common origin for gesture and speech. There are several reasons for this.

The first reason has its roots in the question: Why do gestural phrases align with intermediate phrases? I believe because they are both manifestations of an idea unit, as Kendon suggested, or the unpacking of a growth point, as McNeill proposed. A new idea is expressed, and the units for its expression, in the modalities of gesture and intonation, are a g-phrase and an intermediate phrase, respectively. As discussed, Kendon chose tone units as intonation's package for expressing an idea unit; I have refined this to be Pierrehumbert's smaller entity, the intermediate phrase.

Yerian (1995) noted the lack of a one-to-one correspondence between tone units and gestural phrases. She proposed that the relationship between intonation and gesture was therefore indirect, mediated by idea units. The two modalities were related only as each is related to an idea unit. I believe I've found a closer correspondence, in looking at intermediate phrases, instead of tone units. Gestural phrases are typically (but not always) aligned with a single intermediate phrase, making the relationship between the two modalities more direct.

However, I haven't found a perfect one-to-one relationship. Not every g-phrase in my data aligns with an intermediate phrase. Of those that do, sometimes more than one aligns with an intermediate phrase, or sometimes part of the g-phrase (e.g. the post-stroke hold) doesn't align. And several times a g-phrase appears to span several intermediate phrases, although these cases could be interpreted differently. Nevertheless, I agree with both Kendon and Yerian that idea units are the base of the relationship between gestural phrases and intermediate phrases. In the cases where multiple g-phrases occur within a single intermediate phrase—at syntactic and/or pause boundaries—I believe the idea unit is expressed within the syntactic or pause-defined

chunk. Similarly, in the cases where part of the g-phrase—e.g. the post-stroke hold—doesn't align with an intermediate phrase, I believe the idea unit is expressed within the segment only up to the stroke. Both gestural and intermediate phrases are complex constructions, made up of smaller parts, and perhaps the composite of those parts doesn't always exactly express the idea unit. This complexity also makes annotation difficult, both practically and theoretically. Is a post-stroke hold considered part of a g-phrase merely for convenient packaging? Is a pause with no intonational correlate really not some sort of unit boundary?

To my knowledge, the constructs of a g-phrase and an intermediate phrase were not devised solely with the purpose of capturing an idea unit. Each was constructed with other criteria in mind as well. Kendon, while noting that the obligatory stroke was the essence of the gesture, included the surrounding movement infrastructure—preparation, hold, retraction—to define a g-phrase. Pierrehumbert observed intonational issues in English, such as the domain of downstep, the range of post-nuclear tone spreading, and a variety of contour "levels" to be explained (e.g. high-rise, plateau, low), and realized that the existence of an intermediate phrase, delimited by a high or low phrase accent, was a good solution to these issues. But neither theoretician devised their respective phrase solely around the idea of an idea unit.

The point I'm making is that the theoretical constructs of g-phrase and intermediate phrase may not always and exactly match the extent of idea unit expression. However, I believe that they are good approximations. And the package of idea unit expression is probably never larger than a g-phase (for gesture) or an intermediate phrase (for intonation).

McNeill proposed "triangulating" the growth point within the unfolding discourse, by examining both the stroke of the gesture and the co-occurring word, as these contain the seeds of their growth point source. Growth points—instants in time—are unpacked into idea units—intervals in time—which express the idea born in the growth point. Just as McNeill proposed

triangulating the growth point with the stroke and the co-occurring word, so might we in a similar spirit triangulate the larger idea unit, by looking at the larger intervals of g-phrase and intermediate phrase. I have taken some pains to explain that while these phrases are good approximations of an idea unit, they are not perfect, in that they don't always align perfectly. However, as I've also explained, when they don't align perfectly, there are invariably sub-parts of one phrase which do align neatly with the other. Those sub-parts which do align reveal the existence of the idea unit. When a g-phrase's hold doesn't align with the corresponding intermediate phrase, that hold is outside the span of the idea unit. The idea has been expressed, and the hold is not part of it. Similarly, when two g-phrases fit inside a single intermediate phrase, aligned at a syntactic boundary, there are two idea units existing within what intonationally is a single tune. In the typical case where the g-phrase aligns neatly with the intermediate phrase, the idea unit is delineated by both phrases.

My explanation, that the span of idea units can be recognized by those parts of g-phrases and intermediate phrases which neatly align, may require a re-thinking of the amount of information presented in an idea unit. This amount may be smaller than what others consider. For instance, I gave an example where three g-phrases aligned with a single intermediate phrase, in which the "cupboards" speaker says "*and I have books | that I stack together | to hold it*". I've put vertical bars in the preceding italics at g-phrase boundaries, which are also syntactic constituent boundaries. According to my explanation, each vertical bar delineates a single idea unit. There is first an idea of books, then an idea of stacking, then an idea of the stack holding up a cupboard. Others might argue that the idea arising from the growth point is the single larger idea of stacking books to hold a cupboard. But the three gestures are qualitatively different, and I claim that they are different because they each express a different idea, although the ideas are then combined to a larger idea. McNeill would say that the three gestures reveal three different

growth points. And I am saying that the alignments between gesture and intonation will reveal the three idea units unpacking those three growth points.

To sum up my discussion so far, I believe that gesture and intonation are related at the level of g-phrase and intermediate phrase, and that this is because these phrases are typically the level at which idea units are expressed, in the respective modalities. Just as McNeill explored the unpacking of a growth point within gesture and words, so now we have the level of unpacking within the third modality of intonation.

And just as Kendon and McNeill took the phonological synchrony rule (gesture coincides with or slightly precedes the co-occurring word, but never follows it) as evidence that gesture shares a common origin with speech, so now can we provide further evidence. We can extend this rule to a new modality—intonation—and to a new level—g-phrases and intermediate phrases. Gestural phrases typically slightly precede their corresponding intermediate phrases. Again, gesture surfaces first, and is not a by-product of the utterance.

Even in the cases where gestures took place in silence (e.g. placing imaginary cupboard doors in the air), these took place neatly during a pause (between intermediate phrases), because the gestural phrase was expressing an idea unit by itself. They did not partially overlap with speech or intonation; an idea unit is either expressed by one modality singly, or multiple modalities in concert. This is similar to Emmorey's (1999) observation, described earlier, that signers, when gesturing, don't make manual gestures concurrent with signing, but instead alternate signing with non-sign gestures. Emmorey felt these gestures were communicative, depicting an event being described by the signer.

This leads to another point, which I believe lends support to the idea that gesture is communicative. I've just said that an idea unit can be expressed with either just one modality, or with multiple modalities. I've also stressed that some aspects of gesture and intonation relate not

with nearly every gesture, as gesture and words do, but rather in occasional (yet definite) ways. Pragmatic synchrony between intonation and gesture occurs intermittently, yet quite clearly. And the rhythms of each modality often both diverge and converge. The point I'd like to make, as have others before me, is that a speaker has various modalities at their disposal, and for any given idea, they may choose one or more modalities to express that idea. Words are the default modality, of course, but gesture and intonation are frequently deployed. Neither is required, which is why they don't always surface to carry the message. Other methods for expressing an idea may include non-intonational aspects of prosody (such as loudness, pausing, or silence), different aspects of body movement (such as gaze, facial movements, or torso shifts), changes or emphases in rhythm which involve all of the above, and even pragmatic tools such as shared knowledge, the environment, or Gricean implicatures. As Starkey Duncan put it: "No single communication modality is required in order to display a signal" (Duncan 1972, p. 291).

Gesture is also communicative because it patterns with the communicative modality of intonation. Although pitch can be paralinguistic, expressing qualities such as the speaker's emotion, age and sex, pitch can also be meaningful. Much of the work done on intonation has been to equate intonational patterns with meaning. An L+H* pitch accent indicates contrast; when a reversal-gesture as described above coincides exactly with this pitch accent, it communicates the sense of contrast as well.

In terms of gesture theory, then, I believe gesture's relationship with intonation lends support to the communicative function of gesture, and to the fact that gesture shares a common origin with speech. I don't deny that gesture may have a production-aiding function, however, as there is too much evidence on that side of the debate to be discounted. It may be that gesture serves both functions, as others have proposed. But there is no way that gesture is *not* communicative, whatever other functions it may serve.

Now let's turn to the implications of my findings for intonation theory. Here, the implication is simple. Since intermediate phrases have a gestural correlate among native speakers of English, then there is independent evidence for their existence in English.

Again, the basis here is cognitive, in that an idea unit underlies both gestural and intermediate phrases. It seems intuitive that an intermediate phrase typically presents a single idea. In all the debate about the domain of downstep, or about the tonal qualities of a phrase accent in conjunction with a boundary tone, one can lose sight of the fact that an intermediate phrase is a real entity, serving a real purpose: to express a single idea. The gestural counterpart of an intermediate phrase, then, is not just a convenient piece of independent evidence, but rather a deeper explanation of why an intermediate phrase would exist in the first place.

I've based my discussion on the concept of an *idea unit*. Can this be defined in a clear and falsifiable way? Kendon, who introduced the concept to gesture research, described it in two ways (1988b, p. 264). He first mentioned that it is usually marked by intonational contours, e.g. a tone unit. For my purposes, however, such a definition is circular. I'm claiming that idea units manifest as roughly equivalent units in gesture and intonation. The devil's advocate is asking how I define idea units in the first place. It would be circular to define idea units in terms of intonational units, and then claim that they manifest as such.

Therefore, I can't use Kendon's description of idea units in terms of intonational units, since that's what I'm trying to show. Kendon's other description of the idea unit was to equate it with a minimal unit of sense. Intuitively, this seems understandable, but can it be formally defined? What is *sense*? Is it semantic or pragmatic information? What is a *minimal unit* of sense? Is it a content word (e.g. noun, verb, or adjective?) Is it a syntactic constituent? A predicate? An NP plus VP? Must a minimal unit of sense contain an obligatory piece of new information, along with optional old information?

To examine this further, let's take an example. The "drywall" subject produced the following two sentences, in close succession: *It's neat to see the process, though* followed by *Somebody in my neighborhood needed, uh, she wanted some walls down*. The first sentence was accompanied by a single intermediate phrase, and also by a single g-phrase (in which the hand metaphorically traced out steps in a process). My line of reasoning would have the minimal unit of sense here be an entire sentence, about how it's neat to see a process. In the next sentence, the word *somebody* has its own intermediate phrase and its own g-phrase. This is true also for the prepositional phrase *in my neighborhood*. Do we now say that the minimal units of sense are far smaller than a sentence? Why did the speaker choose to make her minimal unit of sense be a sentence in the first case, a noun in the second, and a prepositional phrase in the third? Is it because in the sentence the only piece of sense is the concept of "process" (with the rest of the sentence merely supporting that concept)? Was it important to the speaker to have *somebody* be a separate idea from *in my neighborhood*? Could the speaker in another situation have produced a combined *somebody in my neighborhood*, spanned by a single intermediate phrase and g-phrase?

The point is that we can't know objectively what's in the speaker's mind, without resorting to mind-reading. It's fine to theoretically propose a minimal unit of sense, but if a speaker is free to adjust the size of that unit based on what she'd like to convey next, it undermines a testable definition. Given this reality, what can be done?

It's interesting to see how McNeill handled another psychological abstraction: the growth point, which could be considered the seed of what becomes expressed as an idea unit. The growth point, like the larger idea unit, is defined theoretically. It is "the utterance's primitive stage, the earliest form of the utterance in deep time, and the opening up of a microgenetic process that yields the surface utterance form as the final stage" (1992, p. 220). This is clearly (by current research methods) an untestable theoretical definition, as a researcher cannot "see" the

growth point with any current instrumentation. McNeill chose instead to *infer* (or "triangulate") the growth point from the surface utterance and accompanying gesture. More specifically, the initial idea contained in the growth point is inferred from the gestural stroke, and its accompanying word(s).

Inferring a psychological abstraction, then, may be the best we can do, until cognitive science provides tools to know exactly what idea a speaker has in her brain. It's not an entirely satisfactory answer, as it introduces the risk of subjectivity on the part of the researcher when working with these concepts.

Therefore, in my discussions above, I'll have to stick with Kendon's intuitive, but not formally definable, definition of an idea unit as a minimal unit of sense. And corollary to McNeill's inferring the growth point from the gestural stroke and accompanying word(s), I propose that we can infer the larger idea unit from the larger g-phrase and intermediate phrase.

Apropos this topic, Susan Duncan's comment (personal communication, February 25, 2004), on my proposed correlation of g-phrases and intermediate phrases, was that I'm approximating what she, McNeill, and colleagues call a "speech-gesture production pulse" (Susan Duncan 2002); a burst of speech with accompanying gesture which roughly manifests an idea being expressed. She emphasized, however, that such a pulse is not reliably definable in terms of any units from any modality.

I've discussed the implications of my findings for both gesture theory and intonation theory, grounding them in deeper cognitive processes. Do my findings have other implications that reach below an integrated surface expression? Perhaps the temporal and rhythmic integration of the two modalities does. I started by noting that, on average, tones cluster around gestural events, regardless of type. I then refined the phonological synchrony rule to show the tight coupling of apexes of movement and pitch accents. And I discovered that the two channels are

related rhythmically. The hands may set a faster tempo than the head due to their greater agility, and speech may be rhythmically slower still due to pulmonary constraints, yet the differing tempos interrelate in complex yet clear ways. This relationship extends to eyeblinks, and even to other conversational participants. A plausible explanation for this extensive and subconscious rhythmic relationship is that all these human processes are driven by an underlying oscillatory mechanism. Many others have proposed this, as discussed earlier. In fact, Loritz (2002) argues that language is rhythmic because our brains are evolutionarily wired to be rhythmic, from our phylogenetic roots as an undulating flagellum, to a vertebrate with a beating heart, to a bipedally walking homo sapiens.

To summarize this discussion, gesture and intonation are related in terms of timing, structure and meaning. This relationship is based on three underlying causes. First, gesture and intonation are two means available to speakers for communicating an idea. Second, the two modalities have a common origin, in which the idea to be communicated is unpacked into a gestural phrase and an intermediate phrase. Third, the two channels are both driven by the same underlying human rhythm.

7 Conclusion and Future Work

This dissertation has investigated the relationship between gesture and intonation. I studied digital videos of four subjects conversing freely with friends. I annotated the videos for intonation, using the ToBI framework, and for gesture, using the guidelines published by McNeill and colleagues. Analyzing the thousands of time-stamped annotations both statistically and by eye, I explored five questions.

First, do body movement and pitch rise and fall together, to reflect increased and decreased emotional tension? I found no evidence for this.

Second, each modality has hypothesized units. Do any of these units align between the modalities? I found that apexes of body movement aligned with pitch accents, and gestural phrases aligned with intermediate phrases. The latter alignment reflects the expression of an idea unit in the two modalities. I also discovered that tones, regardless of type, clustered on average near gestural annotations, regardless of whether those annotations were the start, end, or sole time point of a gestural unit.

Third, do the various unit types within each modality correlate? I found no correlations.

Fourth, do the various meanings contributed by each modality correlate? I found a rich variety of pragmatic reinforcement, in which the two channels served the same pragmatic function simultaneously. This correlation did not happen with every gesture, but only occasionally, as intonation has only a limited repertoire of meaning compared to gesture.

Fifth, how are the modalities related rhythmically? I found that though hands, head, and speech each have their own range of tempos, the tempos often interrelated, diverging, converging, and synchronizing on strong rhythmic points. The rhythmic relationship extended to eyeblinks, and between conversational participants.

In sum, gesture and intonation are related in terms of timing, structure, and meaning. I believe the correlation is not a surface phenomenon, but rather stems from underlying cognitive and rhythmic processes. The relationship lends support to a theory of a common origin of gesture and speech, and also provides independent evidence for the existence of intermediate phrases in English.

This dissertation has been an initial exploration, using relatively new technology (digital video annotation) in a relatively young field. There are many horizons towards which future work could advance, and I'm eager to see others join the endeavor. I'll list some possibilities for future research, beginning with a down-to-earth, close-at-hand topic: annotation.

First of all, for annotation of gesture and intonation, more data is better data. Annotation is so time-consuming that we really have very little data, compared for example to the text annotation community. One solution is simply to have more researchers doing work in this field, each annotating more data to incrementally add to the community's collection at large. Yet while independent researchers have to date added to our collective *knowledge* of the subject, there has been little sharing of actual annotated *data*.

Part of this is due to problems inherent in sharing multimodal data, as discussed in Loehr and Harper (2003). Digital video files take up huge amounts of storage space, and are difficult to share. Even when the files are compressed, researchers must take care to use common compression schemes, else the files are unreadable. And subject privacy must be respected.

Not only is it difficult to share the data, it's also difficult to share the annotations, because there is no single standard for gesture annotation. The intonation field is in much better shape. ToBI is an agreed-upon standard, with published guidelines, tutorials, and inter-annotator agreement figures. The gesture field would do well to follow suit.

My first recommendation for future work, therefore, would be to address the mundane yet important problem of providing an infrastructure for the gesture community to effectively share data. This, in turn, rests upon the ability of the community to agree upon a standard for annotation. This goal might be too idealistic; body movement has a possibly infinite number of phenomena to be annotated. Yet perhaps some subset of body movement could be codified.

With more annotations, and with community-reviewed annotations, a clearer picture might emerge out of this complex field. In my own research, I saw several trends which might have become significant with more data. For example, adaptors might have patterned differently with tones than "true" gestures, which would be a reasonable finding.

Having discussed the need for more annotations in general, I'll also suggest several modifications to the annotation guidelines I used. As I discussed, although I found ToBI an excellent framework to use, there is the problem that edge tones are placed at the end of words, rather than at the end of voicing. It would be interesting to code these tones at the end of voicing, and see if any correlations with gesture became clearer. Another suggestion requires no change to ToBI at all, but rather in my use of it. I only annotated break indexes of levels 3 and 4, as they corresponded to the intonational units I was interested in: intermediate and intonational phrases. But, as discussed, gestural phrases were sometimes aligned with pausal or syntactic breaks within intermediate phrases. It would be worth annotating all of ToBI's break index levels, and see if any of the lower levels correlate with g-phrases (and hence, idea unit expression).

The gestural hierarchy outlined by Kendon and McNeill might do with some experimental tweaking, as well. As mentioned, often post-stroke holds and retractions seemed to have a different status when correlating gesture phrases to intermediate phrases. It would be interesting to annotate these as existing outside of g-phrases, and see if a clearer picture emerged in correlation with intonation, or with any other phenomenon, for that matter.

I'll now turn to perhaps more interesting ideas for future work. My findings were based on adult native speakers of American English, in free-flowing conversations with same-sex friends. Would other parameters yield different results? How would the gesture/intonation relationship differ based on different speaker tasks, such as giving directions, or re-telling a story, or solving a problem? If the participants were strangers, would the findings be different? Did one of my subjects blink on the stressed syllables of her conversational partner because the two were good friends, who unconsciously knew each other's rhythms?

Individuals have idiosyncratic manners of speaking, and the same is true for gesturing. The "sous-chef" subject, for instance, overlaid beats on holds more than the other three subjects. What exactly are the differences between individuals in gesture, intonation, and their interaction? Does the speaker's sex make a difference?

I based one of my findings on the concept of an idea unit, which is expressed in both gesture and intonation. Assuming that an idea unit is a universal phenomenon, how is it expressed gesturally and intonationally across different languages? In languages with intermediate phrases, is the relationship similar to what I've claimed for American English? In languages without intermediate phrases, is there a relationship? What does the gesture/intonation relationship look like in tone languages? In sign languages? As discussed, both Wilcox and Liddell have speculated that gradient modulation of signs in sign language is a form of intonation, and Emmorey has noted gesturing interspersed with signs. How are signs, gesture, and intonation related? Regardless of the language, spoken or sign, tone language or not, I would predict that there is a surface manifestation of an idea unit in various modalities, and that these manifestations will somehow align.

The concept of rhythm is fascinating, and definitely worth pursuing. Again, more data will help here. When the tempo changes within a conversation, what's going on? Is it a

discourse segment change? A speaker change? A shift from one predominant modality to another? And can we pin down the complex interaction of tempos which I conveniently labeled jazz music, for lack of a more precise description?

Interactional synchrony has been under-examined. An empirical investigation of rhythm between speakers is needed. How often, for instance, do people blink on their partner's stressed syllables?

Other phenomena also deserve further investigation. I didn't study facial expressions at all. These are certainly related to intonation, and I believe a study of facial gestures in relation to intonation is worth a dissertation by itself. Perhaps Bolinger's parallel hypothesis might be validated after all, when looking at the face. And the face is certainly related rhythmically to the rest of the body and to speech. In fact, the eyebrows alone are worth studying. I once observed a speaker make an entire utterance with only the eyebrows, by raising them after making an important statement and gazing at his audience, as if to say, "What do you think?"

Among all the other things speakers attend to consciously or unconsciously, it's fascinating to watch how they alternate giving attention to, and gazing at, multiple listeners. The subjects in my research appeared to give their two listeners equal time, or roughly equal periods of gaze, and to switch their gaze at pause boundaries, even within intermediate phrases. This is only a casual observation, and is worth checking more carefully. If a topic is more relevant to one listener, how much more attention does the speaker give that listener? Does the speaker then compensate with extra time for the other listener? Do the shifts have to do with discourse segments?

There are endless other topics one could explore in this arena, because the arena is fundamentally human behavior, which is a vast, complex, universe. Wittgenstein (1922/2001) observed, "Everyday language is a part of the human organism and is no less complicated than

it." To study language, then, is to study humans, both culturally and cognitively. And studying language includes studying the movements of the body and the frequency of the larynx. Condon and Ogston (1967) put it succinctly: "Language, in its natural occurrence as speech, is never disembodied but is always manifested through behavior." Two aspects of that behavior are gesture and intonation. They are related, and their relationship is due to an underlying source of human language.

Appendices

Appendix A: Subject Consent Form

Consent Form for Participation in an
Experiment on Human Communication

You have been invited to participate in a research project conducted by Dan Loehr, a PhD student at Georgetown University. This project researches how people communicate.

If you agree to participate, you will be asked to converse freely and naturally with your conversational partner(s), on any topics you wish, for one hour. You will be given a list of suggested topics of conversation, but these are suggestions only.

There are no foreseeable risks or chances of discomfort.

You will be video- and audio-taped. The information gathered during this experiment will be transcribed and may be made available to others, but at no time will your name or an image of your face be released or associated with any data.

Your participation is completely voluntary, and you may edit or delete any audio or video tapes made of you, or withdraw from the study at any time and for any reason. There is no penalty for withdrawing or not participating.

There are no costs to you or any other party. You will not be paid for your participation. The personal benefits for participation include the ability to help advance the cause of research in human communication.

If you have any questions or complaints after this session is over, you may contact Dan Loehr at (703) 883-6765. You may also contact the Georgetown University Institutional Review Board at (202) 687-5594, if you have any questions or comments regarding your rights as a participant in this research.

This project has been reviewed according to Georgetown University procedures governing your participation in this research.

I, _____(print name) state that I have read and understand the above information. I agree to participate in the human communication experiment under the conditions outlined above. I acknowledge that I have been given a copy of this consent form to keep.

Signature _____

Date _____

Appendix B: Suggested List of Conversational Topics

Possible topics for conversation:

1. What should be done about traffic in the Washington area?
2. What do you think about the 2000 presidential election?
3. Is Texas Rangers shortstop Alex Rodriguez worth being paid a quarter of a billion dollars over the next ten years?
4. Who really should have won the Oscars last year?
5. Was the United States right in returning Elian Gonzalez to his father in Cuba?
6. What do you think about the "Survivor" television series?
7. What do you think about cloning animals?
8. What should be done about the energy crisis in California?
9. Is the World Wide Web worthwhile?
10. Seen any good movies lately?

Appendix C: Perl Programming Scripts

Following are three Perl programming scripts used to answer the question: Do the unit types of gesture and intonation correlate?

The scripts do this by going through Anvil's time-stamped exports, and building a table of co-occurrences of movement and tone types within a given number of milliseconds (passed in to the scripts).

Script q3_cleanup.prl cleans up the Anvil export files.

Script q3_array.prl contains the row/column definitions of the output table.

Script q3.prl does the actual counting of tones occurring within a given number of milliseconds of movements. It then prints the output table, as well as a log file for manual inspection and verification of the counting.

```
## Q3_cleanup.prl: take anvil export files and clean them up:
# e.g. change Anvil label "f0e" to appropriate label, e.g. "PA_",
# change commas to periods, round decimals to milliseconds,
# change numbered labels to strings (e.g. 1 -> "up"),
# and get rid of unwanted columns (e.g "obscured")
#
# Input: basename of files to process (without the underscore)
# e.g. 12Jan_Cam3_1340

sub round;

$basename = $ARGV[0];

@g_phrases_labels =
       ('none', 'deictic', 'emblem', 'iconic', 'metaphoric', 'adaptor');

@file_name_list =
       ('H1_beats','H1_phrases','H2_beats','H2_phrases',#'head_phases',
        'PA','iph','IP');

foreach $file_name (@file_name_list) {

       $file_name_label = $file_name . "_";
       $in_file = $basename . "_" . $file_name . ".txt";
       $out_file = $basename . "_" . $file_name . "_clean.txt";

       open (IN_FILE, "$in_file") || die "cannot open $in_file for read";
       open (OUT_FILE, ">$out_file")
              || die "cannot open $out_file for write";

       $first_line = <IN_FILE>; # skip first line

       while (<IN_FILE>) {
              if(($file_name eq "H1_phrases") or ($file_name eq "H2_phrases")) {
                     ($element,$start,$end,$g_phrases) = split (' ', $_, 4);
              }
              elsif (($file_name eq "H1_beats") or ($file_name eq "H2_beats")) {
                     ($element,$start,$end) = split (' ', $_, 3);
              }
              elsif (($file_name eq "PA") or ($file_name eq "iph") or
                            ($file_name eq "IP")) {
                     ($element,$start,$end,$tone) = split;
              }
              else { print "Unknown file_name $file_name\n"; }

              $element =~ s/f0e/$file_name_label/;
              $start =~ s/\,/\./;
              $start = &round ($start, 3);
              $end =~ s/\,/\./;
              $end = &round ($end, 3);

              if(($file_name eq "H1_phrases") or ($file_name eq "H2_phrases")) {
                     $g_phrases = $g_phrases_labels[$g_phrases];
                     print OUT_FILE "$element\t$start\t$end\t$g_phrases\n";
              }
```

```
		elsif (($file_name eq "H1_beats") or ($file_name eq "H2_beats")) {
			$element_type = "beat";
			print OUT_FILE "$element\t$start\t$end\t$element_type\n";
		}
		elsif (($file_name eq "PA") or ($file_name eq "iph") or
				($file_name eq "IP")) {
			# only need end time for tone proper
			print OUT_FILE "$element\t$end\t$tone\n";
		}

	} # end while

	close IN_FILE;
	close OUT_FILE;

} # end foreach loop

################################################################################

# Now cat the mvt files together for the next step, and ditto for int files
$f1 = $basename . "_H1_clean.txt";
$f2 = $basename . "_H2_clean.txt";
#$f3 = $basename . "_head_clean.txt";
$f4 = $basename . "_all_mvt.txt";
$redirect = '\>';
system ("cat",$f1,$f2,$f3,$redirect,$f4);

$f1 = $basename . "_PA_clean.txt";
$f2 = $basename . "_iph_clean.txt";
#$f3 = $basename . "_IP_clean.txt";
$f4 = $basename . "_all_tone.txt";
$redirect = '\>';
system ("cat",$f1,$f2,$f3,$redirect,$f4);

################################################################################

sub round {
	$num = shift (@_);
	$places = shift (@_);
	$point = index ($num, ".");
	# pad input with zeroes out to proper places, plus one for precision
	while (length($num) < $point+$places+2) {$num .= "0";}
	$new_num = substr($num, 0, $point+$places+1);
	$deciding_digit = substr($num, $point+$places+1, 1);
	if ($deciding_digit >= 5) {
		$new_num += 0.001;
		}
	# pad result with 0's out to proper places, in case we rounded up to a 10
	while (length($new_num) < $point+$places+1) {$new_num .= "0";}
	return $new_num;

} # end sub round
```

```
## q3_array.prl: the code to set up table for Question 3:

=for comment

      H* !H* L* L*+H L*+!H L+H* L+!H* H+!H* H- !H- L- H% L% %H noTone toneTotals


beat
iconic
metaphoric
deictic
emblem
adaptor
noMvt
mvtTotals

=cut

@toneList =
       ('H*','!H*','L*','L*+H','L*+!H','L+H*','L+!H*','H+!H*','X*','H-','!H-',
       'L-','X-','H%','L%','%H','X%','noTone','toneTotals');
@mvtList = (
'beat',
'iconic',
'metaphoric',
'deictic',
'emblem',
'adaptor',
'noMvt',
'mvtTotals');

%table3 = (
       beat => {
              "H*" => 0,
              "!H*" => 0,
              "L*" => 0,
              "L*+H" => 0,
              "L*+!H" => 0,
              "L+H*" => 0,
              "L+!H*" => 0,
              "H+!H*" => 0,
              "X*" => 0,
              "H-" => 0,
              "!H-" => 0,
              "L-" => 0,
              "X-" => 0,
              "H%" => 0,
              "L%" => 0,
              "%H" => 0,
              "X%" => 0,
              noTone => 0,
              toneTotals => 0,
       },
       iconic => {
              "H*" => 0,
```

```
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
        "L+!H*" => 0,
        "H+!H*" => 0,
        "X*" => 0,
        "H-" => 0,
        "!H-" => 0,
        "L-" => 0,
        "X-" => 0,
        "H%" => 0,
        "L%" => 0,
        "%H" => 0,
        "X%" => 0,
        noTone => 0,
        toneTotals => 0,
    },
    metaphoric => {
        "H*" => 0,
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
        "L+!H*" => 0,
        "H+!H*" => 0,
        "X*" => 0,
        "H-" => 0,
        "!H-" => 0,
        "L-" => 0,
        "X-" => 0,
        "H%" => 0,
        "L%" => 0,
        "%H" => 0,
        "X%" => 0,
        noTone => 0,
        toneTotals => 0,
    },
    deictic => {
        "H*" => 0,
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
        "L+!H*" => 0,
        "H+!H*" => 0,
        "X*" => 0,
        "H-" => 0,
        "!H-" => 0,
        "L-" => 0,
        "X-" => 0,
        "H%" => 0,
```

```
        "L%" => 0,
        "%H" => 0,
        "X%" => 0,
        noTone => 0,
        toneTotals => 0,
    },
    emblem => {
        "H*" => 0,
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
        "L+!H*" => 0,
        "H+!H*" => 0,
        "X*" => 0,
        "H-" => 0,
        "!H-" => 0,
        "L-" => 0,
        "X-" => 0,
        "H%" => 0,
        "L%" => 0,
        "%H" => 0,
        "X%" => 0,
        noTone => 0,
        toneTotals => 0,
    },
    adaptor => {
        "H*" => 0,
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
        "L+!H*" => 0,
        "H+!H*" => 0,
        "X*" => 0,
        "H-" => 0,
        "!H-" => 0,
        "L-" => 0,
        "X-" => 0,
        "H%" => 0,
        "L%" => 0,
        "%H" => 0,
        "X%" => 0,
        noTone => 0,
        toneTotals => 0,
    },
    noMvt => {
        "H*" => 0,
        "!H*" => 0,
        "L*" => 0,
        "L*+H" => 0,
        "L*+!H" => 0,
        "L+H*" => 0,
```

```
                "L+!H*" => 0,
                "H+!H*" => 0,
                "X*" => 0,
                "H-" => 0,
                "!H-" => 0,
                "L-" => 0,
                "X-" => 0,
                "H%" => 0,
                "L%" => 0,
                "%H" => 0,
                "X%" => 0,
                noTone => 0,
                toneTotals => 0,
        },
        mvtTotals => {
                "H*" => 0,
                "!H*" => 0,
                "L*" => 0,
                "L*+H" => 0,
                "L*+!H" => 0,
                "L+H*" => 0,
                "L+!H*" => 0,
                "H+!H*" => 0,
                "X*" => 0,
                "H-" => 0,
                "!H-" => 0,
                "L-" => 0,
                "X-" => 0,
                "H%" => 0,
                "L%" => 0,
                "%H" => 0,
                "X%" => 0,
                noTone => 0,
                toneTotals => 0,
        },
);
```

```
## Q3.prl: take cleaned-up anvil export files and produce Table to answer
## question 3: do the unit types of gesture and intonation correlate?

=for comment

      H* !H* L* L*+H L*+!H L+H* L+!H* H+!H* H- !H- L- H% L% %H noTone toneTotals


beat
iconic
metaphoric
deictic
emblem
adaptor
#headMvt
noMvt
mvtTotals

=cut

sub overlap;

require ("q3_array.prl");

# Input: basename of files to process (without the underscore)
# e.g. 12Jan_Cam3_1340
# and msec interval to check for overlap (e.g. 100)

if ($#ARGV != 1) {
       print "\nUsage: perl q3.prl basename overlap_in_msec\n\n";
       print "e.g. perl q3.prl 12Jan_Cam3_1340 100\n\n";
       exit;
}

$basename = $ARGV[0];
$overlap_interval = $ARGV[1];

# Loop through entries in movement file
# For each entry in the movement file, check each entry in int file for overlap
#      if overlap, increment appropriate cell in table
#      if no overlap, increment appropriate "noTone" cell in table

$mvt_file = $basename . "_all_mvt.txt";
$tone_file = $basename . "_all_tone.txt";
$log_file = $basename . "_log.txt";
$table_file = $basename . "_table.txt";

open (MVT_FILE, "$mvt_file") || die "cannot open $mvt_file for read";
open (LOG_FILE, ">$log_file") || die "cannot open $log_file for write";
open (TABLE_FILE, ">$table_file") || die "cannot open $table_file for write";
```

```
while (<MVT_FILE>) {

	($mvt_element,$mvt_start_time,$mvt_end_time,$mvt_element_type) = split;

	# Check for overlap with intonation
	open (TONE_FILE, "$tone_file") || die "cannot open $tone_file for read";

	$at_least_one_overlapping_tone = 0;

	while (<TONE_FILE>) {

		($tone_element,$tone_time,$tone_element_type) = split;

		if (&overlap ($mvt_start_time, $mvt_end_time, $tone_time,
			$overlap_interval)) {
			$at_least_one_overlapping_tone = 1;
			$table1{$mvt_element_type}{$tone_element_type}++;
			print LOG_FILE "$mvt_element ($mvt_element_type,
				$mvt_start_time, $mvt_end_time) & $tone_element
				($tone_element_type, $tone_time)
				occur within $overlap_interval msec\n";
		} # end if overlap

	} # end while loop through TONE_FILE

	close TONE_FILE;

	#if no overlap, then the mvt had no tone with it
	if (!$at_least_one_overlapping_tone) {
		$table1{$mvt_element_type}{noTone}++;
		print LOG_FILE "$mvt_element ($mvt_element_type, $mvt_start_time,
			$mvt_end_time)
			has no overlapping tone within $overlap_interval msec\n";
	}

} # end while loop through MVT_FILE

close MVT_FILE;

# Next, need to check if any tones occurred without movement
# So loop through entries in intonation file
# For each entry in intonation file, check each entry in movement file for
overlap
#	if overlap with _any_ of the movement elements, do nothing -
#		this coincidence has already been counted
#	if no overlap, increment appropriate "noMvt" cell in table

open (TONE_FILE, "$tone_file") || die "cannot open $tone_file for read";
```

```
while (<TONE_FILE>) {

	($tone_element,$tone_time,$tone_element_type) = split;

	open (MVT_FILE, "$mvt_file") || die "cannot open $mvt_file for read";
	$at_least_one_overlapping_mvt = 0;
	while (<MVT_FILE>) {
		($mvt_element,$mvt_start_time,$mvt_end_time,$mvt_element_type)
		= split;
		if (&overlap
		($mvt_start_time, $mvt_end_time, $tone_time, $overlap_interval)) {
			$at_least_one_overlapping_mvt = 1;
		} # end if overlap
	} # end second-pass while loop through MVT_FILE
	close MVT_FILE;

	#if no overlap, then the mvt had no tone with it
	if (!$at_least_one_overlapping_mvt) {
		$table1{noMvt}{$tone_element_type}++;
		print LOG_FILE "$tone_element ($tone_element_type, $tone_time)
			has no overlapping movement within
			$overlap_interval msec\n";
	}

} # end second-pass while loop through TONE_FILE
close TONE_FILE;

# Finally, calculate totals, put in appropriate cells
# sum up the rows (tones)
foreach $movement (@mvtList) {
	$cur_row_total = 0;
	foreach $tone (@toneList) {
		$cur_row_total += $table1{$movement}{$tone};
	}
	$table1{$movement}{toneTotals} = $cur_row_total;
}
# sum up the columns (mvts)
foreach $tone (@toneList) {
	$cur_col_total = 0;
	foreach $movement (@mvtList) {
		$cur_col_total += $table1{$movement}{$tone};
	}
	$table1{mvtTotals}{$tone} = $cur_col_total;
}

# Print out the table
# print top row (tone) labels
print "mvt "; # needed to keep columns lined up
print TABLE_FILE "mvt "; # needed to keep columns lined up
foreach $tone (@toneList) {
	print "$tone ";
	print TABLE_FILE "$tone ";
	}
print "\n";
print TABLE_FILE "\n";
```

```
# print movements for each tone
foreach $movement (@mvtList) {
	print "$movement: ";
	print TABLE_FILE "$movement: ";
	foreach $tone (@toneList) {
		print "$table1{$movement}{$tone} ";
		print TABLE_FILE "$table1{$movement}{$tone} ";
	}
	print "\n";
	print TABLE_FILE "\n";
}

close LOG_FILE;
close TABLE_FILE;

##############################################################################

sub overlap {

	$answer = 0;

	$m_start = shift (@_);
	$m_end = shift (@_);
	$t_time = shift (@_);
	$window = shift (@_);

	$window = $window/1000; # passed in as whole msec, make it be a decimal

	# tone start is irrelevant
	# tone end (passed in as t_time) is the point in time of the tone

	if (($t_time >= $m_start - $window) &&
		($t_time <= $m_end + $window)) {
		$answer = 1;
	}

	return $answer;

} # end sub overlap
```

References

Aboudan, R. & Geoffrey B. (1996). Cross-cultural similarities in gestures: The deep relationship between gestures and speech which transcends language barriers. *Semiotica, 111*(3/4), 269-294.

Bacon, F. (2000). *Advancement of Learning* (M. Kiernan, Ed.). Oxford University Press. (Original work published 1605)

Bavelas, J., & Chovil, N. (1995). Gestures specialized for dialogue. *Personality and Social Psychology Bulletin, 21*(4), 394-405.

Bavelas, J., Chovil, N., Lawrie, D., & Wade, W. (1992). Interactive gestures. *Discourse Processes 15*, 469-489.

Beattie, G., & Aboudan, R. (1994). Gestures, pauses and speech: An experimental investigation of the effects of changing social context on their precise temporal relationships. *Semiotica, 99*(3/4) 239-272.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90*, 35-56.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*(4), 438-462.

Beattie, G., & Shovelton, H. (2003) *An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach*. Manuscript submitted for publication.

Beckman, M., & Elam, G. A. (1997). *Guidelines for ToBI labeling, version 3*. Retrieved November 10, 2003 from http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in English and Japanese. *Phonology Yearbook, 3*, 255-310.

Benner, A. (2001, June). *The onset of gestures: General and contextual effects for different categories of gesture in spontaneous narratives.* Presentation at Orage (Orality and Gesture), Aix-en-Province, France.

Birdwhistell, R. (1952). *Introduction to kinesics: (An annotation system for analysis of body motion and gesture)*. Louisville, KY: University of Louisville.

Birdwhistell, R. (1970). *Kinesics and context*. Philadelphia: University of Pennsylvania Press.

Bloomfield, L. (1933). *Language*. Holt, Rinehart, & Winston.

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

Bolinger, D. (1982). Nondeclaratives from an intonational standpoint. In R. Schneider, K. Tuite, & R. Chametzky (Eds.), *Papers from the Parasession on Nondeclaratives.* Chicago: Chicago Linguistic Society.

Bolinger, D. (1983). Intonation and gesture. *American Speech, 58*(2), 156-174.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford, CA: Stanford University Press.

Bulwer, J. (1974). *Chirologia: Or the natural language of the hand and Chironomia: Or the art of manual rhetoric* (J. Cleary, Ed.). Carbondale, IL: Southern Illinois University Press. (Original work published 1644)

Butcher, C. (1994). *The onset of gesture-speech integration: When hand and mouth move together.* Doctoral Dissertation, University of Chicago.

Butterworth, B., & Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In R. N. Campbell & P. T. Smith (Eds.), *Recent advances in the psychology of language: Formal and experimental approaches* (pp. 347-360). New York: Plenum Press.

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review, 96*, 168-174.

Cassell, J., McNeill, D., & McCullough, K-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition, 7*(1).

Clark, H. H. (1996). *Using Language.* Cambridge University Press.

Cleary, J. (1974). Editor's Introduction. In Bulwer, J. (1974). *Chirologia: Or the natural language of the hand and Chironomia: Or the art of manual rhetoric* (J. Cleary, Ed.). Carbondale, IL: Southern Illinois University Press. (Original work published 1644)

Cohen, A. (1977). The Communicative Function of Hand Illustrators. *Journal of Communication, 27*, 54-63.

Condillac, E. B. (1973) *Essai sur l'origine des connaissances humaines.* Auvers-sur-Oise, France: Editions Galilée. (Original work published 1746)

Condillac, E. B. (1982) Logic, or the first developments of the art of thinking. In *Philosophical writings of Etienne Bonnot, Abbé de Condillac* (F. Philip & H. Lane, Trans.). Hillsdale, NJ: Lawrence Erlbaum. (Original work published 1792)

Condon, W. (1964). *Process in communication*. Unpublished manuscript, Western Psychiatric Institute and Clinic, Pittsburgh PA.

Condon, W. (1976). *An analysis of behavioral organization*. In W. Stokoe & H. R. Bernard (Eds.), *Sign Language Studies 13*, Silver Spring, MD: Linstok Press.

Condon, W., & Ogston, W. (1966). Soundfilm analysis of normal and pathological behavior patterns. *Journal of Nervous and Mental Disorders, 143*, 338-47

Condon, W., & Ogston, W. (1967). A segmentation of behavior. *Journal of Psychiatric Research, 5*, 221-235

Cosnier, J. (1982). Communications et langages gestuels. In J. Cosnier, A. Berrendoner, J. Coulon, & C. Orecchioni (Eds.), *Les voies du langage* (255-304). Paris: Dunod.

Creider, C. (1986). Interlanguage comparisons in the study of the interactional use of gesture. *Semiotica, 62*, 147-163.

Creider, C. (1978). Intonation, Tone Group and Body Motion in Luo Conversation. *Anthropological Linguistics, 20*(7), 327-339.

Cruttenden, A. (1986). *Intonation*. Cambridge University Press.

Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge University Press.

Crystal, D., & Quirk, R. (1964). *Systems of prosodic and paralinguistic features in English*. The Hague: Mouton.

Darwin, C. (1998). *The Expression of the Emotions in Man and Animals* (3rd ed.) *(*P. Ekman, Ed.). Oxford University Press. (Original work published in 1872).

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture*. Cambridge University Press.

Dittmann, A. T., & Llewellyn, L. G. (1969). Body movement and speech rhythm in social conversation. *Journal of Personality and Social Psychology, 11*(2), 98-106.

Duncan, Starkey (1972). Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology, 23*, 283-292.

Duncan, Starkey, & Fiske, D. (1977). *Face to face interaction: Research, methods, and theory.* Hillsdale, NJ: Lawrence Erlbaum.

Duncan, Susan (1996). *Grammatical form and 'thinking-for-speaking' in Mandarin Chinese and English: An analysis based on speech-accompanying gestures*. Doctoral Dissertation, University of Chicago.

Duncan, Susan (2002). *Procedure for developing a gesture-annotated speech transcript based on multiple passes through a single narration*. Unpublished manuscript, University of Chicago.

Efron, D. (1941). *Gesture and Environment*. Morningside Heights, NY: King's Crown Press. Republished 1972 as *Gesture, Race, and Culture*. The Hague: Mouton.

Ekman, P. (1998). Introduction to the third edition. In Darwin, C. (1998). *The Expression of the emotions in man and animals* (3rd ed.) *(*P. Ekman, Ed.). Oxford University Press. (Original work published in 1872).

Ekman, P. (1999). Emotional and conversational nonverbal signals. In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (pp. 44-55). Oxford University Press.

Ekman, P., Friesen, W., & Scherer, K. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica, 16*, 23-27.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49-98.

Emmorey, K. (1999). Do signers gesture? In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (pp. 133-159). Oxford University Press.

Erickson, F. (1981). Money tree, lasagna bush, salt and pepper: Social construction of topical cohesion in a conversation among Italian-Americans. In D. Tannen (Ed*.), Analyzing discourse: Text and talk*. Washington, DC: Georgetown University Press.

Erickson, F., & Shultz, J. (1982). *The counselor as gatekeeper: Social interaction in interviews*. New York: Academic Press.

Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language, 36*, 13-33.

Feyereisen, P. (1999). Neuropsychology of communicative movements. In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (pp. 3-25). Oxford University Press.

Feyereisen, P., & de Lannoy, J.-D. (1991). *Gestures and speech: Psychological investigations.* Cambridge University Press.

Freedman, N., & Hoffman, S. P. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills, 24*, 527-539.

Gill, S. P., Kawamori, M., Katagiri, Y., & Shimojima, A. (1999). Pragmatics of body moves. *Proceedings of the Third International Cognitive Technology Conference, San Francisco.*

Goffman, E. (1983). The interaction order. *American Sociological Review, 48*, 1-17. Reprinted in C. Lemert, & A. Branaman (Eds.), *The Goffman reader*, Oxford, England: Blackwell.

Goldin-Meadow, S. (1997). When gestures and words speak differently. *Current Directions in Psychological Science, 6*(5), 138-143.

Goldin-Meadow, S. (2000). Beyond words: The importance of gesture to researchers and learners. *Child Development, 71*(1), 231-239.

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science, 2*(1), 67-74.

Goldman-Eisler, F. (1967). Sequential temporal patterns and cognitive processes in speech. *Language and Speech, 10*, 122-32.

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica, 62*(1/2), 51-75.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of the Second International Conference on Spoken Language Processing (429-432), Banff, Canada.*

Halliday, M. A. K. (1985). *Introduction to functional grammar*. London: Edward Arnold.

Henderson, A., Goldman-Eisler, F., & Skarbek, A. (1966). Sequential temporal patterns in spontaneous speech. *Language and Speech, 9*, 207-216.

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the Association of Computational Linguistics.*

Hobbs, J. (1990). The Pierrehumbert-Hirschberg theory of intonational meaning made simple: Comments on Pierrehumbert and Hirschberg. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication* (313-323). Boston, MA: MIT Press.

Inkelas, S. & Zec, D. (1995). Syntax-phonology interface. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (535-549). Blackwell.

Iverson, J. M. & Goldin-Meadow, S. (1999). What's communication got to do with it? Gesture in children blind since birth. *Developmental Psychology, 33*(3), 453-467.

Iverson, J. M., Tencer, H. L., Lany, J., & Goldin-Meadow, S. (2000). The relation between gesture and speech in congenitally blind and sighted language-learners. *Journal of Nonverbal Behavior, 24*(2), 105-130.

Kelly, S. & Barr, D. (1997). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*, 577-592.

Kelso, J. A. S., Tuller, B. H., & Harris, K. S. (1983). *A 'dynamic pattern' perspective on the control and coordination of movement.* In P. MacNeilage (Ed.), *The production of speech* (pp. 137-173). New York: Springer-Verlag.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. Siegman & B. Pope (Eds.), *Studies in dyadic communication*. New York: Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key (Ed.), *The Relationship of Verbal and Nonverbal Communication* (207-227). The Hague: Mouton.

Kendon, A. (1982). The study of gesture: Some remarks on its history. *Recherches Sémiotiques/Semiotic Inquiry, 2*, 45-62.

Kendon, A. (1983). Gesture and speech: How they interact. In J. M. Wiemann & R. P. Harrison (Eds.), *Nonverbal Interaction*. Beverly Hills, CA: Sage Publications.

Kendon, A. (1988a). Goffman's approach to face-to-face interaction. In P. Drew & A. Wootton (Eds.), *Erving Goffman: Exploring the interaction order*. Boston: Northeastern University Press.

Kendon, A. (1988b). *Sign languages of aboriginal Australia: Cultural, semiotic and communicative perspectives.* Cambridge University Press.

Kendon, A. (1993). Human gesture. In K. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution*. Cambridge University Press.

Kendon, A. (1996). An agenda for gesture studies. *Semiotic Review of Books, 7*(3), 8-12.

Kettebekov, S., Yeasin, M., Krahnstoever, N., Sharma, R. (2002). Prosody based co-analysis of deictic gestures and speech in weather narration broadcast. *Proceedings of the third international conference on language resources and evaluation (LREC 2002), Multimodal resources and multimodal systems evaluation*.

Kita, S. (2000). how representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

Kita, S., van Gijn, I., & van der Hulst, H. (1997). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Proceedings of Gesture and sign language in human-computer interaction, International gesture workshop, Bielefeld, Germany.*

Kipp, M. (2001). Anvil—a generic annotation tool for multimodal dialogue. *Proceedings of the 7th European conference on speech communication and technology (Eurospeech)* (1367-1370), Aalborg, Denmark.

Krauss, R. M., Chen, Y. & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

Krauss, R. M., Morrel-Samuels, P. & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61*, 743-754.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge University Press.

Levelt, W., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24*, 133-164.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Loehr, D. (2001, March). *Intonation, gesture, and discourse*. Paper presented at the Georgetown University Round Table on Languages and Linguistics.

Loehr, D., & Harper, L. (2003). Commonplace tools for commonplace interactions: Practitioners' notes on entry-level video analysis. *Visual Communication, 2*(2), 225-233.

Loritz, D. (2002). *How the brain evolved language*. Oxford University Press.

Mayberry, R. & Jaques, J. (2000). Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In D. McNeill (Ed.), *Language and Gesture*. Cambridge University Press.

McClave, E. (1991). *Intonation and gesture*. Doctoral Dissertation, Georgetown University, Washington DC.

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

McMahon, A. (2003). Phonology and the holy grail. *Lingua Franca, 113*, 103-115.

McNeill, D. (1985). So you think gestures are nonverbal*? Psychological Review, 92*(3), 350-371.

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

McNeill, D. (1997). Growth points cross-linguistically. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization*. Cambridge University Press.

McNeill, D. (1999). Triangulating the growth point. In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (77-92). Oxford University Press.

McNeill, D. (2000a). Introduction. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

McNeill, D. (2000b). *Catchments and contexts: Non-modular factors in speech and gesture production*. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

McNeill, D. (2000c). *Growth points and catchments*. Retrieved March 22, 2000 from http://vislab.cs.wright.edu/KDI/workshop2000/McNeill%20Presentation%20folder/index.htm

McNeill, D., Cassell, J., & Levy, E. (1993). Abstract deixis. *Semiotica, 95*(1/2), 5-19.

McNeill, D. & Duncan, Susan (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., & Ansari, R. (2001a). Catchments, prosody, and discourse. In *Oralité et gestualité, ORAGE 2001*.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X.-F., & Ansari, R. (2001b). Catchments, prosody, and discourse. *Gesture, 1*, 9-33.

McQuown, N. A. (Ed.) (1971). *The natural history of an interview*. Microfilm collection of manuscripts on cultural anthropology, 15th Series, Joseph Regenstein Library, University of Chicago.

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), 615-622.

Nakatani, C., Grosz, Ahn, B. D., & Hirschberg, J. (1995). *Instructions for annotating discourses.* (Tech. Rep. No. TR-21-95). Boston, MA: Harvard University, Center for Research. in Computer Technology.

Nobe, S. (1996). *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production*. Doctoral Dissertation, University of Chicago.

Nobe, S. (2000). Where do *most* spontaneous representational gestures actually occur with respect to speech? In D. McNeill (Ed.), *Language and gesture*. Cambridge University Press.

Philbin, R., Gelman, M., & Gifford, K. (Executive Producers). (2000, January 12). *Live with Regis and Kathie Lee* [Television broadcast]. Burbank: Buena Vista Television.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Doctoral Dissertation, Massachusetts Institute of Technology, Boston MA.

Pierrehumbert, J., & Beckman, M. (1988) *Japanese tone structure*. Boston, MA: MIT Press.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication* (271-311). Boston, MA : MIT Press.

Pike, K. (1967). *Language in relation to a unified theory of the structure of human behavior*. The Hague: Mouton.

Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, *Yokohama, Japan*.

Pittenger, R., Hockett, C., & Danehy, D. (1960). *The first five minutes: A sample of microscopic interview analysis*. Ithaca, NY: Paul Martineau.

Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K.-E., Furuyama, N., & Ansari, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000), 2*, 247-254.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.-E., & Ansari, R. (2001). Gesture and speech multimodal conversational interaction. *VISLab Report: VISLab-01-01*. Retrieved August 16, 2001, from http://vislab.cs.wright.edu/

Quintillianus, M. F. (1979). *Institutio Oratio* (H. E. Butler, Trans.). London: William Heinemann.

Rimé, B. (1982). The elimination of visible behavior from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology, 12*, 113-129.

Rimé, B. (1983). Nonverbal communication or nonverbal behavior? Towards a cognitive-motor theory of nonverbal behavior. In W. Doise & S. Moscovici (Eds.), *Current issues in European social psychology, 1*, 85-141. Cambridge University Press.

Rimé, B. & Schiaratura, L. (1991). Gesture and speech. In R. S. Feldman & B. Rimé. (Eds.), *Fundamentals of Nonverbal Behavior*. Cambridge University Press.

Sapir, E. (1974). The unconscious patterning of behavior in society. In B. Blount (Ed.), *Language, culture, and society: A book of readings.* Cambridge, MA: Winthrop. (Original work published 1927)

Scheflen, A. E. (1964). The significance of posture in communication systems. *Psychiatry, 27*, 316-331.

Scheflen, A. E. (1968). Human communication: Behavioral programs and their integration in interaction. *Behavioral Sciences, 13*, 44-55.

Schegloff, E. (1984). On some gestures' relation to speech. In. J. M. Atkinson & J. Heritage (Eds.), *Structure of social action: Studies in conversational analysis* (266-296). Cambridge University Press.

Scherer, K. & Wallbott, H. (1985). Analysis of nonverbal behavior. In T. van Dijk (Ed.), *Handbook of discourse analysis*. New York: Academic Press.

Scollon, R. (1981a). *Tempo, density, and silence: Rhythms in ordinary talk.* University of Alaska, Fairbanks, Center for Cross-Cultural Studies.

Scollon, R. (1981b). The rhythmic integration of ordinary talk. In D. Tannen (Ed.), *Analyzing discourse: Text and talk. Georgetown University Round Table on Languages and Linguistics 1981* (335-349). Washington, DC: Georgetown University Press.

Shukla, S. (1996). Śikşa:s, pra:tiśa:khyas, and the vedic accent. In K. R. Jankowsky (Ed.), *Multiple perspectives on the historical dimensions of language* (269-279). Münster, Germany: Nodus Publikationen.

Kjölander, K., & Beskow, J. (2000). WaveSurfer – An open source speech tool. *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China.*

Silverman, K., Beckman M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J.. (1992). TOBI: A standard for labeling English prosody. *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Banff, Canada.*

Slobin, D. I. (1987). Thinking for speaking. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Papers from the 13th Annual Meeting of the Berkeley Linguistics Society* (480-519). Berkeley, CA: Berkeley Linguistics Society.

Trager, G. L. & Smith, H. L., Jr. (1957). *An outline of English structure.* Washington, DC: American Council of Learned Societies.

Tuite, K. (1993). The production of gesture. *Semiotica, 93*, 83-105.

Valbonesi, L,, Ansari, R., McNeill, D., Quek, F., Duncan, S., McCullough, K.-E., & Bryll, R. (2002). *Temporal correlations of speech and gestures focal points.* Manuscript submitted for publication.

Wittgenstein, L. (2001). *Tractatus Logico-Philosophicus* (D. F. Pears & B. F. McGuinness, Trans.). London: Routledge. (Original work published 1922, New York: Harcourt Brace)

Wolff, C. (1972) *A Psychology of gesture* (A. Tennant, Trans.). New York: Arno Press. (Original work published 1945, London: Methuen)

Wundt, W. (1973) *The Language of Gestures* (J. S. Thayer, C. M. Greenleaf, & M. D. Silberman, Trans.). The Hague: Mouton. (From *Völkerpsychologie*, originally published 1921, Verlag)

Yerian, K. (1995). *Directions in intonation and gesture research: A Review*. Unpublished manuscript, Georgetown University, Washington DC.

Yerian, K. (2000). *The discursive body: Vocal and non-vocal interactional strategies and the strategic construction of (gendered) stances in women's self-defense courses*. Doctoral Dissertation, Georgetown University, Washington DC.

Zipf, G. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston, MA: Houghton Mifflin.