Multimodal Human Discourse: Gesture and Speech

FRANCIS QUEK
Wright State University
DAVID MCNEILL
University of Chicago
ROBERT BRYLL
Wright State University
SUSAN DUNCAN
Wright State University
University of Chicago
XIN-FENG MA
University of Illinois at Chicago
CEMIL KIRBAS
Wright State University
KARL E. MCCULLOUGH
University of Chicago
and
RASHID ANSARI
University of Illinois at Chicago

Gesture and speech combine to form a rich basis for human conversational interaction. To exploit these modalities in HCI, we need to understand the interplay between them and the way in which they support communication. We propose a framework for the gesture research done to date, and present our work on the cross-modal cues for discourse segmentation in free-form gesticulation accompanying speech in natural conversation as a new paradigm for such multimodal interaction. The basis for this integration is the psycholinguistic concept of the coequal generation of gesture and speech from the same semantic intent. We present a detailed case study of a gesture and speech elicitation experiment in which a subject describes her living space to an interlocutor. We perform

The research and preparation for this article was supported by the US National Science Foundation STIMULATE program, Grant No. IRI-9618887, "Gesture, Speech, and Gaze in Discourse Segmentation," and the National Science Foundation KDI program, Grant No. BCS-9980054, "Cross-Modal Analysis of Signal and Sense: Multimedia Corpora and Tools for Gesture, Speech, and Gaze Research." Additional support from Silicon Graphics in the form of an Equipment Grant is acknowledged.

Authors' addresses: F. Quek, R. Bryll, Department of Computer Science and Engineering, Vision Interfaces and Systems Lab, Wright State University, 3640 Colonel Glenn Hwy., Dayton, OH 45435-0001; email: {quek,bryll}@cs.wright.edu.

two independent sets of analyses on the video and audio data: video and audio analysis to extract segmentation cues, and expert transcription of the speech and gesture data by microanalyzing the videotape using a frame-accurate videoplayer to correlate the speech with the gestural entities. We compare the results of both analyses to identify the cues accessible in the gestural and audio data that correlate well with the expert psycholinguistic analysis. We show that "handedness" and the kind of symmetry in two-handed gestures provide effective supersegmental discourse cues.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*theory and methods*; *interaction styles*

General Terms: Languages

Additional Key Words and Phrases: Multimodal interaction, conversational interaction, gesture, speech, discourse, human interaction models, gesture analysis

1. INTRODUCTION

Natural human conversation is a rich interaction among multiple verbal and nonverbal channels. Gestures are an important part of human conversational interaction [McNeill 1992] and they have been studied extensively in recent years in efforts to build computer-human interfaces that go beyond traditional input devices such as keyboard and mouse manipulations. In a broader sense "gesture" includes not only hand movements, but also facial expressions and gaze shifts. For human-computer interaction to approach the level of transparency of interhuman discourse, we need to understand the phenomenology of conversational interaction and the kinds of extractable features that can aid in its comprehension. In fact, we believe that we need new paradigms for looking at this nexus of multimodal interaction before we can properly exploit the human facility with this rich interplay of verbal and nonverbal behavior in an instrumental fashion.

Although gesture may be extended to include head and eye gestures, facial gestures, and body motion, we explore only the relationship of hand gestures and speech in this article. In a case study of 32 seconds of video of a "living-space description" discourse, we demonstrate the efficacy of handedness and symmetry analyses in identifying the shifts in discourse topic. In the course of our presentation we hope to motivate a view of this multimodal interaction that gives us a glimpse into the construction of discourse itself.

1.1 Manipulation and Semaphores

There is a growing body of literature on the instrumental comprehension of human gestures. Predominantly, research efforts are clustered around two kinds of gestures: manipulative and semaphoric. We define manipulative gestures as those whose intended purpose is to control some entity by applying a tight relationship between the actual movements of the gesturing hand/arm with the entity being manipulated. Semaphores are systems of signaling using flags, lights, or arms [Brittanica.com]. By extension, we define semaphoric gestures to be any gesturing system that employs a stylized dictionary of static or dynamic hand or arm gestures.

Research employing the manipulative gesture paradigm may be thought of as following the seminal "Put-That-There" work by Bolt [1980, 1982].

Since then, there have been a plethora of systems that implement finger tracking/pointing, a variety of "finger flying" style navigation in virtual spaces or direct-manipulation interfaces, control of appliances, in computer games, and robot control. In a sense the hand is the ultimate multipurpose tool, and manipulation represents properly a large proportion of human hand use. We have observed, however, that gestures used in communication/conversation differ from manipulative gestures in several significant ways [Quek 1995, 1996]. First, because the intent of the latter is for manipulation, there is no guarantee that the salient features of the hands are visible. Second, the dynamics of hand movement in manipulative gestures differ significantly from conversational gestures. Third, manipulative gestures may typically be aided by visual, tactile, or force-feedback from the object (virtual or real) being manipulated, whereas conversational gestures are typically performed without such constraints. Gesture and manipulation are clearly different entities sharing between them possibly only the feature that both may utilize the same body parts.

Semaphoric gestures are typified by the application of a recognition-based approach to identify some gesture $g_i \in \mathcal{G}$ where $\mathcal{G}$ is a set of predefined gestures. Semaphoric approaches may be termed "communicative" in that gestures serve as a universe of symbols to be communicated to the machine. A pragmatic distinction between semaphoric gestures and manipulative ones is that the semaphores typically do not require the feedback control (e.g., hand-eye, force-feedback, or haptic) necessary for manipulation. Systems operating under this paradigm typically define a set of stylized gesture and head movement symbols that are then recognized by a variety of techniques, including graph labeling [Triesch and von der Malsburg 1996], principal components analysis [Lanitis et al. 1995], hidden Markov models [Yamato et al. 1992; Hofmann et al. 1997; Schlenzig et al. 1994], and neural networks [Schlenzig et al. 1994; Edwards 1997]. Unfortunately such semaphoric handuse is a miniscule percentage of typical handuse in communication.

Both manipulative and semaphoric gesture models suffer significant shortcomings. Although manipulation represents a significant proportion of human natural hand use, natural manipulation situations almost always involve the handling of the artifact being manipulated. Free-hand manipulation interfaces, however, lack such feedback and rely almost exclusively on visual feedback.

Semaphores represent a miniscule portion of the use of the hands in natural human communication. In reviewing the challenges to automatic gesture recognition, Wexelblat [1997] emphasizes the need for development of systems able to recognize natural, nonposed, and nondiscrete gestures. Wexelblat disqualifies systems recognizing artificial, posed, and discrete gestures as unnecessary and superficial:

> If users must make one fixed gesture to, for example, move forward in a system then stop, then make another gesture to move backward, I find myself wondering why the system designers bother with gesture in the first place. Why not simply give the person keys to press: one for forward and one for backward?

He considers the natural gestural interaction to be the only one "real" and useful mode of interfacing with computer systems:

> . . . one of the major points of gesture modes of operation is their naturalness. If you take away that advantage, it is hard to see why the user benefits from a gestural interface at all.

He underscores the need for systems working with truly conversational gestures, and also emphasizes the tight connection of gestures and speech (conversational gestures cannot be analyzed without considering speech). He expresses urgent need for standard datasets that could be used for testing of gesture recognition algorithms. One of his conclusions, however, is that the need for conversational gesture recognition still remains to be proven (by proving, e.g., that natural gesture recognition can improve speech recognition):

> An even broader challenge in multimodal interaction is the question of whether or not gesture serves any measurable useful function, particularly in the presence of speech.

In their review of gesture recognition systems, Pavlović et al. [1997] conclude that natural, conversational gesture interfaces are still in their infancy. They state that most current works "address a very narrow group of applications: mostly symbolic commands based on hand postures or 3D-mouse type of pointing," and that "real-time interaction based on 3D hand model-based gesture analysis is yet to be demonstrated."

1.2 Gesture-Speech Approaches

Several researchers have looked at the speech-gesture nexus. Wexelblat [1995] describes research whose goal is to "understand and encapsulate gestural interaction in such a way that gesticulation can be treated as a datatype—like graphics and speech—and incorporated into any computerized environment where it is appropriate." The author does not make any distinction between the communicative aspect of gesture and the manipulative use of the hand, citing the act of grasping a virtual doorknob and twisting as a natural gesture for opening a door in a virtual environment. The paper describes a set of experiments for determining the characteristics of human gesticulation accompanying the description of video clips subjects have viewed. These experiments were rather naive given the large body of literature on narration of video episodes [McNeill 1992]. The author states that "in general we could not predict *what* users would gesture about," and that "there were things in common between subjects that were not being seen at a full-gesture analysis level. Gesture command languages generally operate only at a whole gesture level, usually by matching the user's gesture to a pre-stored template. . . . [A]ttempting to do gesture recognition solely by template matching would quickly lead to a proliferation of templates and would miss essential commonalities" [of real gestures]. As shown later, this affirms our assertion concerning the characteristics of human gesticulation accompanying speech. Proceeding from these observations, the author describes a system that divides the gesture analysis process into two phases: feature analysis and interpretation ("where meaning is assigned

to the input"). The system uses CyberGloves,[1] body position sensors and eye trackers as inputs (the data are sampled at 20 Hz). The data are then segmented by feature detectors and the features temporally integrated to form "frames" representing phases of a gesture. Beside stating that the outputs of these analyses would go to domain-dependent gesture interpreters that may be built, the system makes no attempt to recognize the gestures. The architecture of the system is similar to the architecture presented in Boehme et al. [1997]: the signals are analyzed by layers of parallel detectors/analyzing units, and the level of abstraction increases with each level of analysis.

Wilson et al. [1996] proposed a triphasic gesture segmenter that expects all gestures to be a rest, transition, stroke, transition, rest sequence. They use an image-difference approach along with a finite-state machine to detect these motion sequences. Natural gestures are, however, seldom clearly triphasic in the sense of this article. Speakers do not normally terminate each gesture sequence with their hands in the rest positions. Instead, retractions from the preceding gesture often merge with the preparation of the next.

Kahn et al. [1994] describe their Perseus architecture that recognizes a standing human form pointing at various predefined artifacts (e.g., Coke™ cans). They use an object-oriented representation scheme with a "feature map" comprising intensity, edge, motion, disparity, and color features to describe objects (standing person and pointing targets) in the scene. Their system reasons with these objects to determine the pointed-at object. Extending Perseus, Franklin et al. [1996] describe an extension of this work to direct and interact with a mobile robot.

There is a class of systems that applies a combination of semaphoric and manipulative gestures within a single system. This class is typified by Pavlovic et al. [1996] and combines HMM-based gesture semaphores (move forward, backward), static hand poses (grasp, release, drop, etc.), and pointing gestures (fingertip tracking using two orthogonally oriented cameras: top and side). The system is used to manipulate graphical DNA models.

Sowa and Wachsmuth [2000, 1999] describe a study based on a system for using coverbal iconic gestures for describing objects in the performance of an assembly task in a virtual environment. They use a pair of CyberGloves for gesture capture, three Ascension Flock of Birds electromagnetic trackers[2] mounted to the subject's back for torso tracking and wrists, and a headphone-mounted microphone for speech capture. In this work, subjects describe contents of a set of five virtual parts (e.g., screws and bars) that are presented to them in wall-size display. The gestures were annotated using the *Hamburg Notation System* for sign languages [Prillwitz et al. 1989]. The authors found that "such gestures convey geometric attributes by abstraction from the complete shape. Spatial extensions in different dimensions and roundness constitute the dominant 'basic' attributes in [their] corpus . . . geometrical attributes can be expressed in several ways using combinations of movement trajectories, hand distances, hand apertures, palm orientations, hand-shapes and index finger direction." In essence,

[1]See www.virtex.com.

[2]See www.ascension-tech.com.

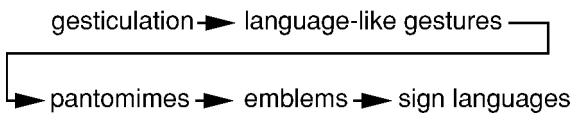

Fig. 1. A continuum of gestures (reproduced from McNeill [1992]).

even with the limited scope of their experiment in which the imagery of the subjects was guided by a wall-size visual display, a panoply of iconics relating to some (hard-to-predict) attributes of each of the five target objects were produced by the subjects.

There is a class of gestures that sits between pure manipulation and natural gesticulation (see Section 2). This class of gestures, broadly termed deictics or pointing gestures, has some of the flavor of manipulation in its capacity of immediate spatial reference. Deictics also facilitate the "concretization" of abstract or distant entities in discourse, and so are the subject of much study in psychology and linguistics. Following Bolt [1980, 1982], work done in the area of integrating direct manipulation with natural language and speech has shown some promise in such combination. Earlier work by Cohen et al [1989, 1998] involved the combination of the use of a pointing device and typed natural language to resolve anaphoric references. By constraining the space of possible referents by menu enumeration, the deictic component of direct manipulation was used to augment the natural language interpretation. Neal et al [1989, 1998] describe similar work employing mouse pointing for deixis and spoken and typed speech in a system for querying geographical databases. Oviatt et al. [Oviatt et al. 1999; Oviatt and Cohen 2000; Oviatt 1999] extended this research direction by combining speech and natural language processing and pen-based gestures. We have argued that pen-based gestures retain some of the temporal coherence with speech as with natural gesticulation [Quek et al. 2000], and this cotemporality was employed in Oviatt et al. [1999], Oviatt and Cohen [2000], and Oviatt [1999] to support mutual disambiguation of the multimodal channels and the issuing of spatial commands to a map interface. Koons et al. [1993, 1998] describe a system for integrating deictic gestures, speech, and eye gaze to manipulate spatial objects on a map. Employing a tracked glove, they extracted the gross motions of the hand to determine such elements as "attack" (motion toward the gesture space over the map), "sweep" (side-to-side motion), and "end reference space" (the terminal position of the hand motion). They relate these spatial gestural references to the gaze direction on the display, and to speech to perform a series of "pick-and-place" operations. This body of research differs from that reported in this article in that we address more free-flowing gestures accompanying speech, and are not constrained to the two-dimensional reference to screen or pen-tablet artifacts of pen or mouse gestures.

2. GESTURE AND SPEECH IN NATURAL CONVERSATION

When one considers a communicative channel, a key question is the degree of formalism required. Kendon [Kendon 1986; McNeill 1992] describes a philology of gesture summarized in Figure 1 which is useful in this consideration. At one end of this continuum is *gesticulation* which describes the free-form

gesturing which typically accompanies verbal discourse. At the other end of this continuum are sign languages which are characterized by complete lexical and grammatical specification. In between we have "language-like gestures" in which the speaker inserts a gesture in the place of a syntactic unit during speech; "pantomimes" in which the gesturer creates an iconic and motion facsimile of the referent, and "emblems" which are codified gestural expressions that are not governed by any formal grammar. (The thumbs-up gesture to signify "all is well" is an example of such a gesture.) This article focuses on the gesticulation end of this continuum.

In natural conversation between humans, gesture and speech function together as a coexpressive whole, providing one's interlocutor access to semantic content of the speech act. Psycholinguistic evidence has established the complementary nature of the verbal and nonverbal aspects of human expression [McNeill 1992]. Gesture and speech are not subservient to each other, as though one were an afterthought to enrich or augment the other. Instead, they proceed together from the same "idea units," and at some point bifurcate to the different motor systems that control movement and speech. For this reason, human multimodal communication coheres topically at a level beyond the local syntax structure.

This discourse structure is an expression of the idea units that proceed from human thought. Although the visual form (the kinds of hand shapes, etc.), magnitude (distance of hand excursions), and trajectories (paths along which hands move) may change across cultures and individual styles, underlying governing principles exist for the study of gesture and speech in discourse. Chief among these is the temporal coherence between the modalities at the level of communicative intent. A football coach may say, "We have to throw the ball into the end zone." Depending on the intent of the expression, the speaker may place the speech and gestural stresses on different parts on the expression. If the emphasis is on the act, the stresses may coincide on the verb THROW, and if the emphasis is on the place of the act, the stresses accompany the utterance END ZONE. This temporal coherence is governed by the constants of the underlying neuronal processing that proceed from the nascent idea unit or growth point [McNeill 2000b; McNeill and Duncan 2000]. Furthermore, the utterance in our example makes little sense in isolation. The salience of the emphasis comes into focus only when one considers the content of the preceding discourse. The emphasis on the act THROW may be in contrast to some preceding discussion of the futility of trying to run the ball earlier in the game. The discourse may have been: "They are stacking up an eight-man front to stop the run . . . We have to THROW the ball into the end zone." The accompanying gesture might have taken the form of the hand moving forward in a throwing motion coincident with the prosodic emphasis on THROW. Alternatively, the emphasis on the place specifier END ZONE may be in contrast to other destinations of passing plays. In this case, the discourse may have been: "We have no more time-outs, and there are only 14 seconds left in the game. . . . We have to throw the ball into the END ZONE." The accompanying gesture may be a pointing gesture to some distal point from the speaker indicating the end zone. This will again be cotemporal with the prosodically emphasized utterance END ZONE. Hence,

gesture and speech prosody features may provide access to the management of discourse at the level of its underlying threads of communicative intent.

We present the underlying psycholinguistic model by which gesture and speech entities may be integrated, describe our research method that encompasses processing of the video data and psycholinguistic analysis of the underlying discourse, and present the results of our analysis. Our purpose, in this article, is to motivate a perspective of conversational interaction, and to show that the cues for such interaction are accessible in video data. In this study, our data come from a single video camera, and we consider only gestural motions in the camera's image plane.

3. PSYCHOLINGUISTIC BASIS

Gesture and speech clearly belong to different modalities of expression but they are linked on several levels and work together to present the same semantic idea units. The two modalities are not redundant; they are *coexpressive*, meaning that they arise from a shared semantic source but are able to express different information, overlapping this source in their own ways. A simple example illustrates. In the living-space text we present below, the speaker describes at one point entering a house with the clause, "when you open the doors." At the same time she performs a two-handed antisymmetric gesture in which her hands, upright and palms facing forward, move left to right several times. Gesture and speech arise from the same semantic source but are nonredundant; each modality expresses its own part of the shared constellation. Speech describes an action performed in relation to it, and gesture shows the shape and extent of the doors and that there are two of them rather than one; thus speech and gesture are coexpressive. Since gesture and speech proceed from the same semantic source, one might expect the semantic structure of the resulting discourse to be accessible through both the gestural and speech channels. Our *catchment* concept provides a locus along which gestural entities may be viewed to provide access to this semantic discourse structure.

The catchment is a unifying concept that associates various discourse components [McNeill 2000b, 2000a; McNeil et al. 2001]. A catchment is recognized when gesture features recur in two or more (not necessarily consecutive) gestures. The logic is that the recurrence of imagery in a speaker's thinking will generate recurrent gesture features. Recurrent images suggest a common discourse theme. These gesture features can be detected and the recurring features offer clues to the cohesive linkages in the text with which they co-occur. A catchment is a kind of thread of visuospatial imagery that runs through the discourse to reveal emergent larger discourse units even when the parts of the catchment are separated in time by other thematic material. By discovering the catchments created by a given speaker, we can see what this speaker is combining into larger discourse units: what meanings are regarded as similar or related and grouped together, and what meanings are being put into different catchments or are being isolated, and thus seen by the speaker as having distinct or less-related meanings. By examining interactively shared catchments, we can extend this thematic mapping to the social framework of

Fig. 2. A frame of the video data collected in our gesture elicitation experiment.

the discourse. Gestural catchments have their own distinctive prosodic profiles and gaze control in that gaze redirection is an accompaniment of catchment (re)introduction.

By discovering a given speaker's catchments, we can see what for this speaker goes together into larger discourse units: what meanings are seen as similar or related and grouped together, and what meanings are isolated and thus seen by the speaker as having distinct or less-related meanings.

Consider one of the most basic gesture features, handedness. Gestures can be made with one hand (1H) or two (2H); if 1H, they can be made with the left hand (LH) or the right (RH); if 2H, the hands can move and/or be positioned in mirror images or with one hand taking an active role and the other a more passive "platform" role. Noting groups of gestures that have the same values of handedness can identify catchments. We can add other features such as shape, location in space, and trajectory (curved, straight, spiral, etc.), and consider all of these as also defining possible catchments. A given catchment could, for example, be defined by the recurrent use of the same trajectory and space with variations of hand shapes. This would suggest a larger discourse unit within which meanings are contrasted. Individuals differ in how they link up the world into related and unrelated components, and catchments give us a way of detecting these individual characteristics or cognitive styles.

4. EXPERIMENTAL METHOD

Hand gestures are seen in abundance when humans try to communicate spatial information. In our gesture and speech elicitation experiment, subjects are asked to describe their living quarters to an interlocutor. This conversation is recorded on a Hi-8 tape using a consumer-quality camcorder (a Sony TR-101 for the results presented here). Figure 2 is a frame from the experimental sequence that is presented here. Two independent sets of analyses are performed

on the video and audio data. The first set of analyses entails the processing of the video data to obtain the motion traces of both of the subject's hands. The synchronized audio data are also analyzed to extract the fundamental frequency signal and speech power amplitude (in terms of the RMS value of the audio signal). The second set of analyses entails the expert transcription of the speech and gesture data. This transcription is done by microanalyzing the Hi-8 videotape using a frame-accurate videoplayer to correlate the speech with the gestural entities. We also perform a higher-level analysis using the transcribed text alone. Finally, the results of the psycholinguistic analyses are compared against the features computed in the video and audio data. The purpose of this comparison is to identify the cues accessible in the gestural and audio data that correlate well with the expert psycholinguistic analysis. We discuss each step in turn.

4.1 Extraction of Hand Motion Traces in Video

In the work described here our purpose is to see what cues are afforded by gross hand motion for discourse structuring. Human hand gesture in standard video data poses several processing challenges. First, one cannot assume contiguous motion. A sweep of the hand across the body can span just 0.25 s to 0.5 s. This means that the entire motion is captured in 7 to 15 frames. Depending on camera field-of-view on the subject, interframe displacement can be quite large. This means that dense optical flow methods cannot be used. Second, because of the speed of motion, there is considerable motion blur. Third, the hands tend to occlude each other.

We apply a parallelizable fuzzy image processing approach known as vector coherence mapping (VCM) [Quek and Bryll 1998; Quek et al. 1999] to track the hand motion. VCM is able to apply spatial coherence, momentum (temporal coherence), motion, and skin color constraints in the vector field computation by using a fuzzy-combination strategy, and produces good results for hand gesture tracking. Figure 3 illustrates how VCM applies a spatial coherence constraint (minimizing the directional variance) in vector field computation. Assume three feature points $p_1^t \cdots p_3^t$ at time t (represented by the squares at the top of the figure) move to their new locations (represented by circles) in the next frame. If all three feature points correspond equally to one another, an application of convolution to detect matches (e.g., by an absolute difference correlation, ADC) from each p_i^t would yield correlation maps with three hotspots (shown as $\mathcal{N}_1^t \cdots \mathcal{N}_3^t$ in the middle of Figure 3). If all three correlation maps are normalized and summed, we obtain the vector coherence map (vcm) at the bottom of the figure; the "correct" correlations would reinforce each other, and the chance correlations would not. Hence a simple weighted summation of neighboring correlation maps would yield a vector field that minimizes the local variance in the computed vector field. We can adjust the degree of coherence enforced by adjusting the contributing weights of the neighboring correlation maps as a function of distance of these maps from the point of interest.

A normalized vcm computed for each feature point can be thought of as a likelihood map for the spatial variance-minimizing vector at that point. This

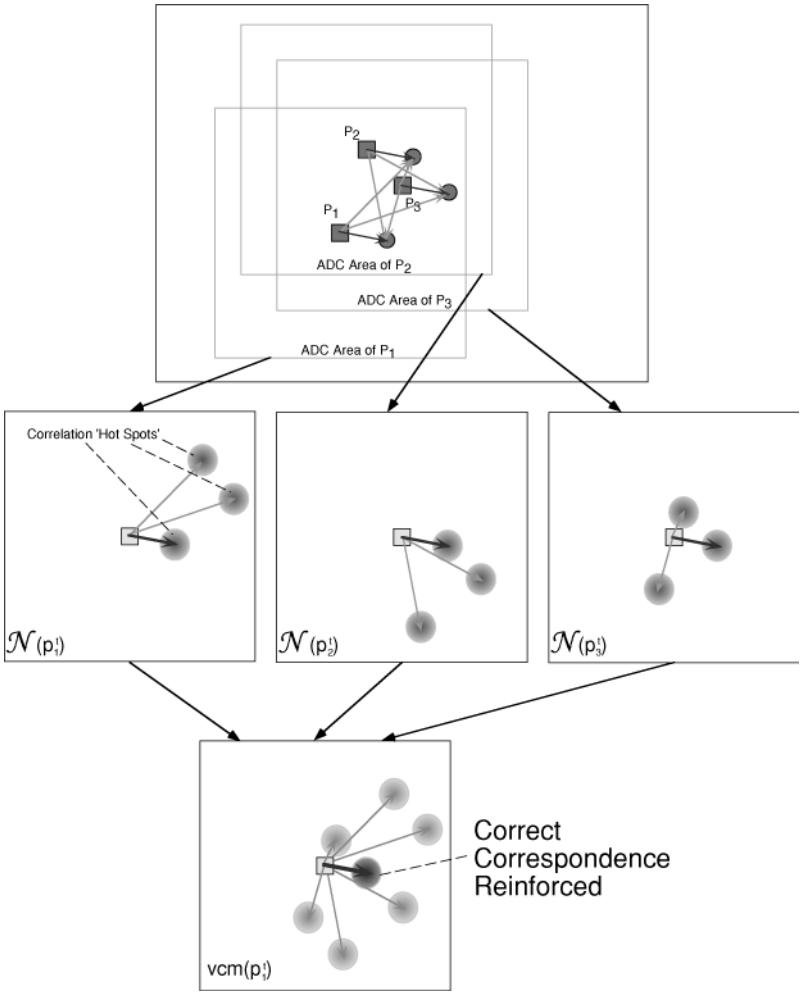

Fig. 3. Spatial coherence constraint in VCM.

may be used in a fuzzy image processing process where other constraints may be fuzzy-ANDed with it. We use a temporal constraint based on momentum, and a skin-color constraint to extract the required hand motion vector fields. Figure 4 shows the effect of incorporating the skin-color constraint to clean up the computed vector fields. This approach is derived in detail in Quek and Bryll [1998] and Quek et al. [1999].

4.2 Audio Processing

Since our experiments require natural interaction between the subject and interlocutor, we decided not to use intrusive head-mounted devices. In the experiments here reported, the audio signal came directly from the built-in microphones on the Sony TR-101 camcorders. Given the irregular distance and orientation between the subject's mouth and the microphone, and other

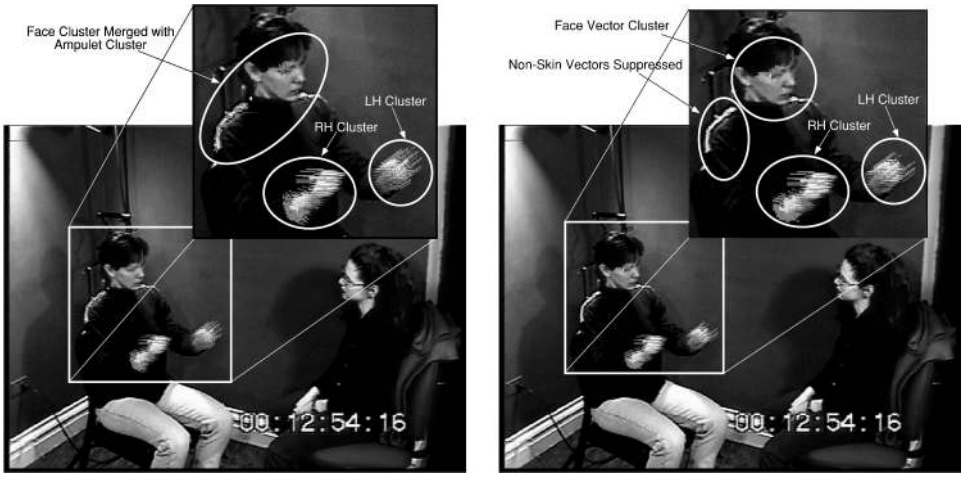

Fig. 4. Hand and head movements tracked in the same sequence without (left) and with (right) the skin color constraint applied.

environmental factors, our analysis has to handle a fair amount of noise. For the experiment reported here, we make use of the speech fundamental frequency F_0 estimate for the speaker to extract voiced speech components, and the amplitude of the speech (in terms of the RMS value of the audio signal). We first obtain an estimate of both the F_0 and RMS of the speech using Entropic's Xwaves+^{TM} software. The resulting F_0 signal contains two kinds of errors: false $F_0 s$ due to background interference, and F_0 harmonics at half and a third of the correct value. For the first kind of error, we apply a median filter to the computed RMS signal, use knowledge about the speaker's typical F_0 to eliminate signals that are too far from this expected value, and suppress low-amplitude F_0 signals if the corresponding signal power (RMS) falls below a threshold. To address the second error component, we apply a running estimate on the speaker's F_0 value. If the F_0 computed by Xwaves+ changes abruptly by a factor of 0.5 or 0.33, it is frequency shifted while preserving the signal amplitude. Finally, since the amplitude of spoken (American) English tends to decline over the course of phrases and utterances [Ladd 1996], we apply a linear declination estimator to group F_0 signal intervals into such declination units.

We compute the pause durations within the F_0 signal. The remaining F_0 signal is mapped onto a parametric trend function to locate likely phrase units. For further discussion of our audio processing algorithm, please see Ansari et al. [1999].

4.3 Detailed Psycholinguistic Analysis

Perceptual analysis of video (analysis by unaided ear and eye) plays an important role in such disciplines as psychology, psycholinguistics, linguistics, anthropology, and neurology. In psycholinguistic analysis of gesture and speech, researchers microanalyze videos of subjects using a high-quality videocassette recorder that has a digital freeze capability down to the specific frame. The

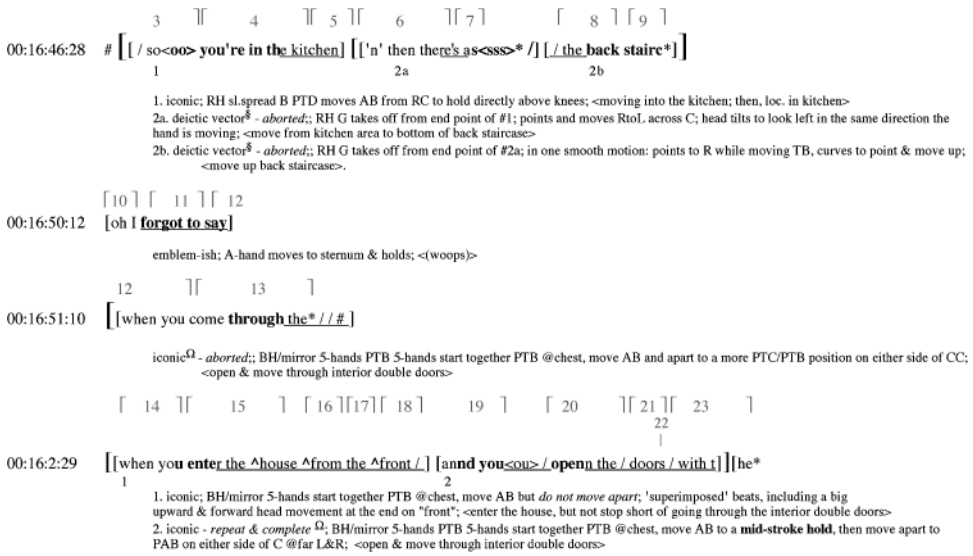

Fig. 5. Sample first page of a psycholinguistic analysis transcript.

analysis typically proceeds in several iterations. First, the speech is carefully transcribed by hand, and typed into a text document. The beginning of each linguistic unit (typically a phrase) is marked by the timestamp of the beginning of the unit on the videotape. Second, the researcher revisits the video and annotates the text, marking co-occurrences of speech and gestural phases (rest-holds, pre-stroke and post-stroke holds, gross hand shape, trajectory of motion, gestural stroke characteristics, etc.). The researcher also inserts locations of audible breath pauses, speech disfluencies, and other salient comments. Third, all these data are formatted onto a final transcript for psycholinguistic analysis. This is a painstaking process that takes a week to 10 days to analyze about a minute of discourse.[3]

The outcome of the psycholinguistic analysis process is a set of detailed transcripts. We have reproduced the first page of the transcript for our example dataset in Figure 5. The gestural phases are annotated and correlated with the text (by underlining, boldface, brackets, symbol insertion, etc.), F_0 units (numbers above the text), and videotape timestamp (time signatures to the left). Comments about gesture details are placed under each transcribed phrase.

In addition to this, we perform a second level-of-discourse analysis based on the transcribed text alone without looking at the video (see Section 5.1.1).

4.4 Comparative Analysis

In our discovery experiments, we analyze the gestural and audio signals in parallel with the perceptual psycholinguistic analysis to find cues for high-level

[3]Since the work reported here, our process is now using speech recognition technology to obtain an approximate time alignment of the transcribed text, and using the Praat Phonetics annotation system [Boersma and Weenik 1996] to obtain millisecond-accurate timestamps. We have also developed a multimedia system to assist in this process [Quek and McNeill 2000; Quek et al. 2001].

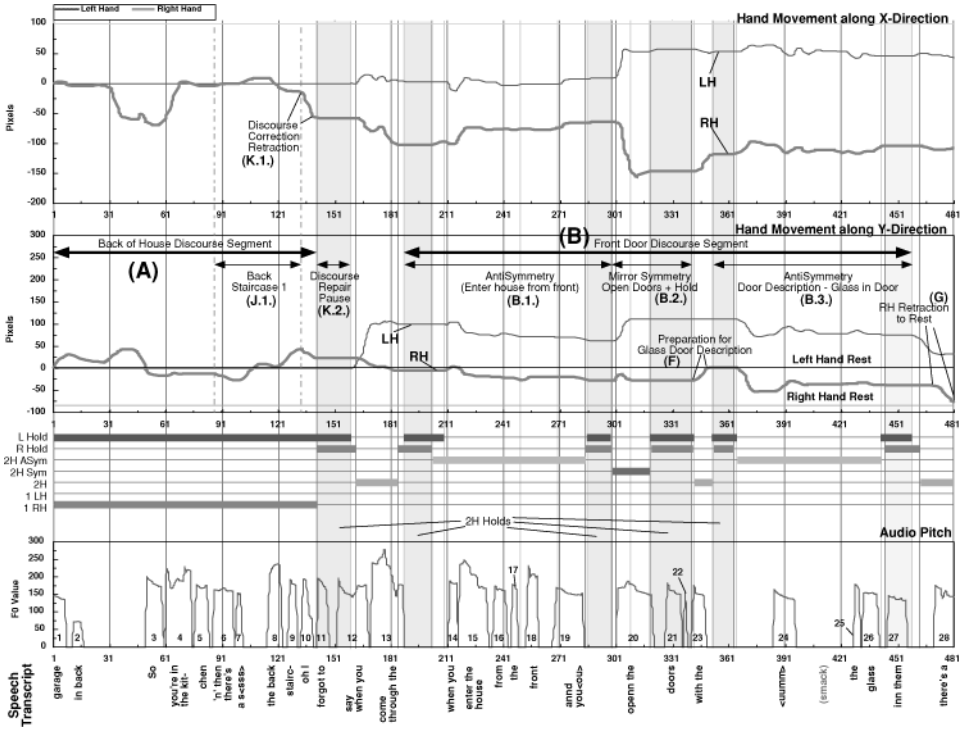

Fig. 6. Hand position, handedness analysis, and F_0 graphs for the frames 1–480.

discourse segmentation that are accessible to computation. The voiced units in the F_0 plots are numbered and related manually to the text. These are plotted together with the gesture traces computed using VCM. First, we perform a perceptual analysis of the data to find features that correlate well between the gesture signal and the high-level speech content. We call this the discovery phase of the analysis. Next, we devise algorithms to extract these features. We present the features and cues we have found in our data.

5. RESULTS

Figures 6 and 7 are summary plots of 32 seconds of experimental data plotted with the videoframe number as the timescale (480 frames in each chart to give a total of 960 frames at 30 fps). The x-axes of these graphs are the frame numbers. The top two plots of each figure describe the horizontal (x) and vertical (y) hand positions, respectively. The horizontal bars under the y-direction plot are an analysis of hand motion features: LH hold, RH hold, 2H antisymmetric, 2H symmetric (mirror symmetry), 2H (no detected symmetry), single LH, and single RH, respectively. Beneath this is the fundamental frequency F_0 plot of the audio signal. The voiced utterances in the F_0 plot are numbered sequentially to facilitate correlation with the speech transcript. We have reproduced the synchronized speech transcript at the bottom of each chart, as it correlates with the F_0 units. The vertical shaded bars that run across the charts mark the

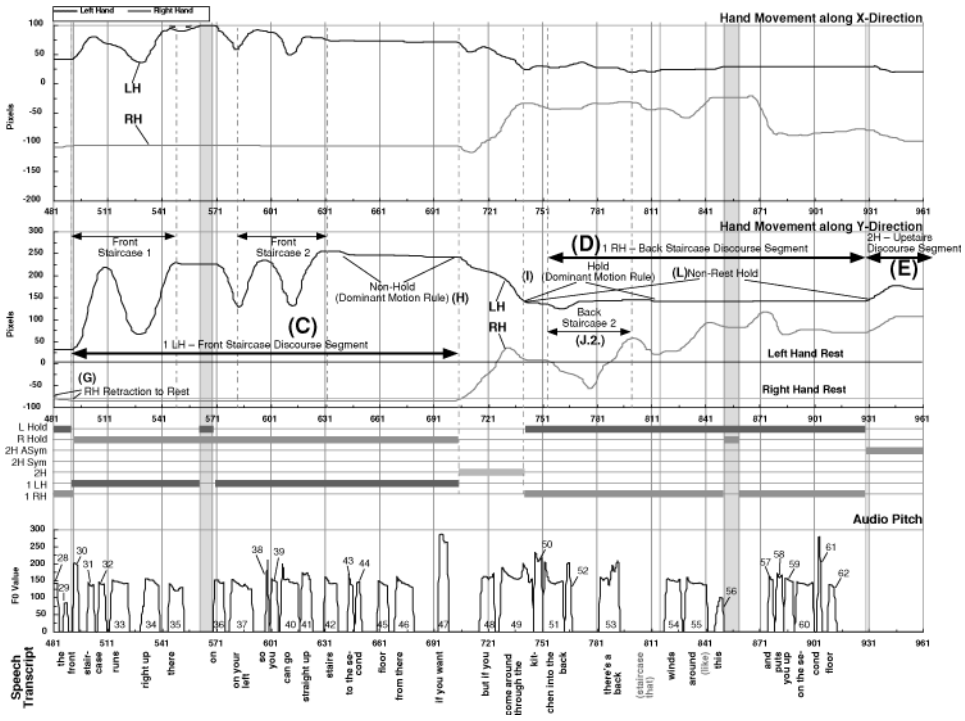

Fig. 7. Hand position, handedness analysis, and F_0 graphs for the frames 481–961.

durations in which both hands are determined to be stationary (holding). We have annotated the hand motion plots with parenthesized markers and brief descriptive labels. We use these labels along with the related frame-number durations for our ensuing discussion.

5.1 Discourse Segments

5.1.1 *Linguistic Segmentation.* Barbara Grosz and colleagues [Nakatani et al. 1995] have devised a systematic procedure for recovering the discourse structure from a transcribed text. The method consists of a set of questions with which to guide analysis and uncover the speaker's goals in producing each successive line of text.

Following this procedure with the text displayed in Figures 6 and 7 produces a three-level discourse structure. This structure can be compared to the discourse segments inferred from the objective motion patterns shown in the gestures. The gesture-based and text-based pictures of the discourse segmentation are independently derived but the resulting correlation is remarkably strong, implying that the gesture features were not generated haphazardly but arose as part of a structured multilevel process of discourse building. Every gesture feature corresponds to a text-based segment. Discrepancies arise where the gesture structure suggests discourse segments that the text-based hierarchy fails to reveal, implying that the gesture modality has captured additional discourse segments that did not find their way into the textual transcription. The

uppermost level of the Grosz-type hierarchy can be labeled "Locating the Back Staircase." It is at this level that the discourse as a whole had its significance. The middle level concerned the first staircase and its location, and the lowest level the front of the house and the restarting of the house tour from there.

5.1.2 *Gesture–Speech Discourse Correlations.* Labels (A) through (E) mark the discourse segments accessible from the gestural traces independently of the speech data. These segments are determined solely from the hold and motion patterns of the speaker's hands. We summarize the correspondences of these segments with the linguistic analysis following.

(A) Back-of-house discourse segment, 1 RH (Frames 1–140): These one-handed gestures, all with the RH, accompany the references to the back of the house that launch the discourse. This 1H catchment is replaced by a series of 2H gestures in (B), marking the shift to a different discourse purpose, that of describing the front of the house. Notice that this catchment feature of 1H–RH gestures (i.e., the LH is holding) reprises itself in segment (D) when the subject returns to describing the back of the house.

(B) Front-door discourse segment, 2 Synchronized Hands (Frames 188–455): Two-handed gestures occur when the discourse theme is the front of the house, but there are several variants and these mark subparts of the theme: the existence of the front door, opening it, and describing it. Each subtheme is initiated by a gesture hold, marking off in gesture the internal divisions of the discourse hierarchy. These subdivisions are not evident in the text and thus not picked up by the text-only analysis that produced the purpose hierarchy and its segmentation (described in Section 5.1.1). This finer-grained segmentation is confirmed by psycholinguistic analysis of the original video.

(B.1) "Enter house from front" discourse segment 2H Antisymmetric (Frames 188–298): Antisymmetric two-handed movements iconically embody the image of the two front doors; the antisymmetric movements themselves contrast with the following mirror-image movements, and convey, not motion as such, but the surface and orientation of the doors.

(B.2) "Open doors" discourse segment 2H Mirror Symmetry (Frames 299–338): In contrast to the preceding two-handed segment, this gesture shows opening the doors and the hands moving apart. This segment terminates in a non-rest two-handed hold of sizeable duration of more than 0.75 s (all other pre-stroke and post-stroke holds are less than 0.5 s in duration). This suggests that it is itself a hold-stroke (i.e., an information-laden component). This corresponds well with the text transcription. The thrusting open of the hands indicates the action of opening the doors (coinciding with the words, "open the"), and the 2H hold-stroke indicates the object of interest (coinciding with the word, "doors"). This agrees with the discourse analysis that carries the thread of the "front door" as the element of focus. Furthermore, the 2H mirror-symmetric motion for opening the doors carries the added information that these are double doors (information unavailable from the text transcription alone).

(B.3) Door description discourse segment 2H Antisymmetric (Frames 351–458): The form of the doors returns as a subtheme in its own right,

and again the movement is antisymmetric, in the plane of the closed doors.

(C) Front staircase discourse segment, 1 LH (Frames 491–704): The LH becomes active in a series of distinctive up–down movements coinciding exactly with the discourse goal of introducing the front staircase. These are clearly one-handed gestures with the RH at the rest position.

(D) Back staircase discourse segment 1 RH (Frames 754–929): The gestures for the back staircase are again made with the RH, but now, in contrast to the (A) catchment, the LH is at a non-rest hold, and still in play from (C). This changes in the final segment of the discourse.

(E) "Upstairs" discourse segment 2H synchronized (Frames 930–): The LH and RH join forces in a final gesture depicting ascent to the second floor via the back staircase. This is another place where gesture reveals a discourse element not recoverable from the text (no text accompanied the gesture).

5.2 Other Gestural Features

Beside the overall gesture hold analysis, this 32 seconds of discourse also contains several examples of gestural features and cues. We have labeled these (F) through (L). The following discussion summarizes the features we found under these labels.

(F) Preparation for glass door description (Frames 340–359): In the middle of the discourse segment on the front door (B), we have the interval marked (F) which appears to break the symmetry. This break is actually the preparation phase of the RH to the non-rest hold (for both hands) section that continues into the strongly antisymmetric (B.3) "glass door" segment. This clarifies the interpretation of the 2H holds preceding and following (F). The former is the post-stroke hold for the "open doors" segment (B.2), and the latter is the pre-stroke hold for segment (B.3). Furthermore, we can then extend segment (B.3) backward to the beginning of (F). This would group F_0 unit 23 ("with the") with (B.3). This matches the discourse segmentation "with the . . . <uumm> glass in them." It is significant to note that the subject has interrupted her speech stream and is searching for the next words to describe what she wants to say. The cohesion of the phrase would be lost in a pure speech pause analysis. We introduce the rule for gesture segment extension to include the last movement to the pre-stroke hold for the segment.

(G) RH retraction to rest (Frames 468–490 straddles both charts): The RH movement labeled (G) spanning both plots terminates in the resting position for the hand. We can therefore extend the starting point of the target rest backward to the start of this retraction for discourse segmentation. Hence, we might actually move the (C) front staircase discourse segment marked by the 1 LH feature backward from frame 491 to frame 470. This matches the discourse analysis from F_0 units 28–30: ". . . there's the front- . . . ." This provides us with the rule of rest-extension to the start of the last motion that results in the rest. The pauseless voiced speech section from F_0 units 28–35 would provide complementary evidence for this segmentation. We have another example for this rule in the LH motion preceding the hold labeled (I), effectively extending

the "back staircase" discourse segment (D) backward to frame 705. Again this corresponds well with the speech transcript.

(H) and (I) Non-Hold for (H) in (C) (Frames 643–704) and Hold for (I) in (D) (Frames 740–811): The LH was judged by the expert coder to be not holding in (H) and it was judged to be holding in (I). An examination of the video shows that in (H) the speaker was making a series of small oscillatory motions (patting motion with her wrist to signify the floor of the "second floor") with a general downward trend. In segment (I), the LH was holding, but the entire body was moving slightly because of the rapid and large movements of the RH. This distinction cannot be made from the motion traces of the LH alone. Here, we introduce a *dominant motion rule* for rest determination. We use the motion energy differential of the movements of both hands to determine if small movements in one hand are interpreted as holds. In segment (H), the RH is at rest, hence any movement in the alternate LH becomes significant. In segment (I), the RH exhibits strong motion, and the effects of the LH motion are attenuated.

(J.1) and (J.2) Back-Staircase Catchment 1 (Frames 86–132), and Back-Staircase Catchment 2 (Frames 753–799): Figure 8 is a side-by-side comparison of the motions of the right hand that constitute the back staircase description. The subject described the spiral staircase with an upward twirling motion of the RH. In the first case, the subject aborted the gesture and speech sequence with an abrupt interruption and went on to describe the front of the house and the front staircase. In the second case, the subject completed the gestural motion and the back staircase discourse. We can make several observations about this pair of gestures that are separated by more than 22 seconds. First, both are gestures of one hand (RH). Second, the general forms of both gestures are similar. Third, up till the discourse repair break, both iterations had exactly the same duration of 47 frames. Fourth, the speaker appears to have already been planning a change in direction in her discourse, and the first motion is muted with respect to the second.

(K.1) and (K.2) Discourse repair retraction (Frames 133–142), and discourse repair pause (Frames 143–159): Segments (K.1) and (K.2) correspond to the speech, "Oh I forgot to say," and flag a repair in the discourse structure. The subject actually pulls her RH back toward herself rapidly and holds an emblematic gesture with an index finger point. Although more data in experiments targeted at discourse repair are needed for us to make a more definitive statement, it is likely that abrupt gestural trajectory changes where the hand is retracted from the gestural space suggest discourse repair. (We are in the process of performing experiments, and building a video-audio-transcription corpus to answer such questions.)

(L) Non-rest hold (Frames 740–929): An interesting phenomenon is seen in the (L) non-rest hold. This is a lengthy hold spanning 190 frames or 6.33 seconds. This means that it cannot be a pre-stroke or post-stroke hold. It could be a hold gesture or a stationary reference hand in a 2H gesture sequence. In the present discourse example it actually serves as an "idea hold." The subject just ended her description of the front staircase with a mention of the second floor. While her LH is holding, she proceeds to describe the back staircase that

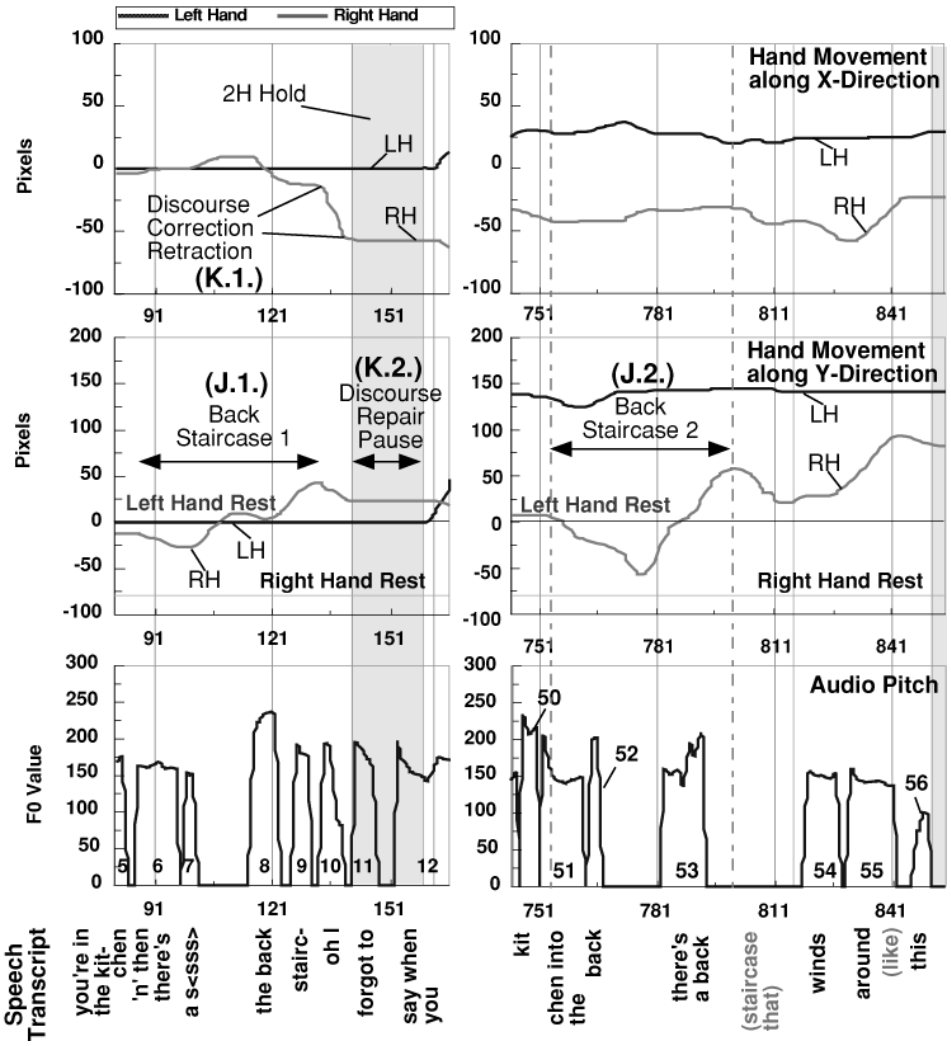

Fig. 8. Side-by-side comparison of the back-staircase catchment.

takes her to the same location. At the end of this non-rest hold, she goes on to a 2H gesture sequence (E) describing the second floor. This means that her discourse plan to proceed to the second floor was already in place at the end of the (C) discourse segment. The LH suspended above her shoulder could be interpreted as holding the upstairs discourse segment in abeyance while she describes the back staircase (segment (D)). Such holds may thus be thought of as supersegmental cues for the overall discourse structure. The non-rest hold (L), in essence, allows us to connect the end of (C) with the (E) discourse segment.

6. DISCUSSION

Natural conversational interaction is a complex composition of multimodal visually and audibly accessible information. We believe that the key to such

interaction is an understanding of the underlying psycholinguistic phenomenology, along with a handle on the kinds of computable visual and audio features that aid its interpretation.

We have shown that conversational discourse analysis using both speech and gesture captured in video is possible. We overviewed our algorithms for gesture tracking in video and showed our results for the extended tracking across 960 video frames. The quality of this tracking permits us to perform the psycholinguistic analysis presented.

In the example discourse analyzed, we have shown strong correlation between handedness and high-level semantic content of the discourse. Where both hands are moving, the kind of synchrony (antisymmetry or mirror symmetry) also provides cues for discourse segmentation. In some cases, the gestural cues reinforce the phrase segmentation based on F_0 pause analysis (segment boundaries coincide with such pauses in the voiced signal). In some other cases the gestural analysis provides complementary information permitting segmentation where no pauses are evident, or grouping phrase units across pauses. The gestural analysis corresponds well with the discourse analysis performed on the text transcripts. In some cases, the gestural stream provides segmentation cues for linguistic phenomena that are inaccessible from the text transcript alone. The finer-grained segmentations in these cases were confirmed when the linguists reviewed the original experimental videotapes.

We also presented other observations about the gestural signal that are useful for discourse analysis. For some of these observations, we have derived rules for extracting the associated gestural features. We have shown that the final motion preceding a pre-stroke hold should be grouped with the discourse unit of that hold. Likewise, the last movement preceding a rest-hold should be grouped with the rest for that hand. When a hand movement is small, our dominant motion rule permits us to distinguish if it constitutes a hold or gestural hand motion. We have also made observations about repeated motion trajectory catchments, discourse interruptions and repair, and non-rest holds.

Our data also show that although 2-D monocular video affords a fair amount of analysis, more could be done with 3-D data from multiple cameras. In the two discourse segments labeled (B.1) and (B.3) in Figure 6, a significant catchment feature is that the hands move in a vertical plane iconic of the surface of the doors. The twirling motion for the back staircase catchment would be more reliably detected with 3-D data. The discourse repair retraction toward the subject's body (K.1) would be evident in 3-D. There is also a fair amount of perspective effects that are difficult to remove from 2-D data. Because of the camera angle (we use an oblique view because subjects are made uncomfortable sitting face-to-face with the interlocutor and directly facing the camera), a horizontal motion appears to have a vertical trajectory component. This trajectory is dependent on the distance of the hand from the subject's body. In our data, we corrected for this effect by assuming that the hand is moving in a fixed plane in front of the subject. This produces some artifacts that are hard to distinguish without access to the three-dimensional data. In the "open doors" stroke in (B.2), for example, hands move in an arc centered around the subject's elbows and shoulders. Hence, there is a large displacement in the "z" direction

(in and away from the body). The LH also appears to move upward and the RH's motion appears to be almost entirely in the x-dimension. Consequently, the LH seems to move a shorter distance than the RH in the x-dimension. This has directed us to move a good portion of our ongoing experiments to 3-D.

7. CONCLUSION

Gesticulation accompanying speech is an exquisite performance bearing imagistic and semantic content. This performance is unwitting, albeit not unintended, by the speaker just as the orchestration of carefully timed exercise of larynx, lips, tongue, lungs, jaw, and facial musculature are brought unwittingly to bear on intended vocal utterances. The work reported here just barely scratches the surface of a compelling scientific direction. This research is necessarily crossdisciplinary, and involves a fair amount of collaboration between the computation sciences and psycholinguistics. The computational requirements for the video processing are significant. On a four-processor ($4 \times$ R10000) Silicon Graphics Onyx workstation, we could process 10 seconds of data in two hours. We have since obtained National Science Foundation support for a supercomputer to handle this data and computationally intensive task. We are in the process of testing the results obtained from the case study reported here on a corpus of experimental video data we are assembling in new human subject discourse elicitations. We are also continuing to determine other cues for discourse analysis in different conversational scenarios through detailed case studies like the one reported here.

The bulk of prior research on gesture interaction has focused on manipulative and semaphoric interfaces. Apart from the use of the hands in direct manipulation, stylized semaphore recognition has dominated communicative gesture research. We contend that far from being naturally communicative, such stylized use of the hand(s) is in fact a rather unnatural means of communication that requires specialized training. Nonetheless, gesture and speech are a potent medium for human-machine interfaces. To tap into the richness of this multimodal interaction, it is essential that we have a better understanding of the "natural" function of gesture and speech. This article introduces a paradigm of analysis that may have applications such as in multimodal speech/discourse recognition, detection of speech repairs in interactive speech systems, and modeling of multimodal elements for distance communication (e.g., with avatar control [Quek et al. 2000]). More information about our ongoing research is available at http://vislab.cs.wright.edu/KDI.

REFERENCES

ANSARI, R., DAI, Y., LOU, J., MCNEILL, D., AND QUEK, F. 1999. Representation of prosodic structure in speech using nonlinear methods. In *Workshop on Nonlinear Signal & Image Processing* (Antalya, TU, June 20–23).

BOEHME, H.-J., BRAKENSIEK, A., BRAUMANN, U.-D., KRABBES, M., AND GROSS, H.-M. 1997. Neural architecture for gesture-based human-machine-interaction. In *Proceedings of the International Gesture Workshop: Gesture & Sign Language in HCI*, I. Wachsmuth and M. Frohlich, Eds., Springer, Bielefeld, Germany, 219–232.

BOERSMA, P. AND WEENIK, D. 1996. Praat, a system for doing phonetics by computer. Tech. Rep. 132, Institute of Phonetic Sciences of the University of Amsterdam.

BOLT, R. A. 1980. Put-that there. *Comput. Graph. 14*, 262–270.

BOLT, R. A. 1982. Eyes at the interface. In *Proceedings of the ACM CHI Human Factors in Computing Systems Conference*, 360–362.

BRITTANICA.COM. Encyclopaedia brittanica web site. http://www.brittanica.com.

COHEN, P., DALRYMPLE, M., MORAN, D., PEREIRA, F., SULLIVAN, J., GARGAN, R., SCHLOSSBERG, J., AND TYLER, S. 1989. Synergistic use of direct manipulation and natural language. In *Human Factors in Computing Systems: CHI'89 Conference Proceedings*, ACM, Addison-Wesley, Reading, Mass., 227–234.

COHEN, P., DALRYMPLE, M., MORAN, D., PEREIRA, F., SULLIVAN, J., GARGAN, R., SCHLOSSBERG, J., AND TYLER, S. 1998. Synergistic use of direct manipulation and natural language. In *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds., Morgan Kaufman, San Francisco, 29–35.

EDWARDS, A. 1997. Progress in sign language recognition. In *Proceedings of the International Gesture Workshop on Gesture & Sign Language in HCI*, I. Wachsmuth and M. Frohlich, Eds., Springer, Bielefeld, Germany, 13–21.

FRANKLIN, D., KAHN, R. E., SWAIN, M. J., AND FIRBY, R. J. 1996. Happy patrons make better tippers creating a robot waiter using Perseus and the animate agent architecture. In *FG96* (Killington, Vt.), 253–258.

HOFMANN, F., HEYER, P., AND HOMMEL, G. 1997. Velocity profile based recognition of dynamic gestures with discrete hidden Markov models. In *Proceedings of the International Gesture Workshop: Gesture & Sign Language in HCI*, I. Wachsmuth and M. Frohlich, Eds., Springer, Bielefeld, Germany, 81–95.

KAHN, R., SWAIN, M., PROJKOPOWICZ, P., AND FIRBY, R. 1994. Gesture recognition using the Perseus architecture. In *Proceedings of the IEEE Conference on CVPR*, IEEE Computer Society, Los Alamitos, Calif., 734–741.

KENDON, A. 1986. Current issues in the study of gesture. In *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, J.-L. Nespoulous, P. Peron, and A. Lecours, Eds., Lawrence Erlbaum, Hillsdale, N.J., 23–47.

KOONS, D., SPARRELL, C., AND THORISSON, K. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*, M. Maybury, Ed., AAAI Press; The MIT Press, Cambridge, Mass., 257–276.

KOONS, D., SPARRELL, C., AND THORISSON, K. 1998. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*, M. Maybury, Ed., AAAI Press; The MIT Press, Cambridge, Mass., 53–62.

LADD, D. 1996. *Intonational Phonology*. Cambridge University Press, Cambridge.

LANITIS, A., TAYLOR, C., COOTES, T., AND AHMED, T. 1995. Automatic interpretation of human faces and hand gestures. In *Proceedings of the International Workshop on Automatic Face & Gesture Recognition* (Zurich) 98–103.

MCNEILL, D. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

MCNEILL, D. 2000a. Catchments and context: Non-modular factors in speech and gesture. In *Language and Gesture*, D. McNeill, Ed., Cambridge University Press, Cambridge, Chapter 15, 312–328.

MCNEILL, D. 2000b. Growth points, catchments, and contexts. *Cogn. Stud. Bull. Japan. Cogn. Sci. Soc. 7*, 1.

MCNEILL, D. AND DUNCAN, S. 2000. Growth points in thinking-for-speaking. In *Language and Gesture*, D. McNeill, Ed., Cambridge University Press, Cambridge, Chapter 7, 141–161.

MCNEILL, D., QUEK, F., MCCULLOUGH, K.-E., DUNCAN, S., FURUYAMA, N., BRYLL, R., MA, X.-F., AND ANSARI, R. 2001. Catchments, prosody and discourse. *Gesture* in press.

NAKATANI, C., GROSZ, B., AHN, D., AND HIRSCHBERG, J. 1995. Instructions for annotating discourses. Tech. Rep. TR-21-95, Center for Research in Computer Technology, Harvard University, Cambridge, Mass.

NEAL, J. G., THIELMAN, C. Y., DOBES, Z., HALLER, S. M., AND SHAPIRO, S. C. 1989. Natural language with integrated deictic and graphic gestures. In *Proceedings of the Speech and Natural Language Workshop* (Cape Cod, Mass.) 410–423.

NEAL, J. G., THIELMAN, C. Y., DOBES, Z., HALLER, S. M., AND SHAPIRO, S. C. 1998. Natural lan-

guage with integrated deictic and graphic gestures. In *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds., Morgan Kaufman, San Francisco, 38–51.

OVIATT, S. 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the CHI 99, Vol. 1*, 576–583.

OVIATT, S. AND COHEN, P. 2000. Multimodal interfaces that process what comes naturally. *Commun. ACM 43*, 3, 43–53.

OVIATT, S., DEANGELI, A., AND KUHN, K. 1999. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the CHI 97, Vol. 1*, 415–422.

PAVLOVIĆ, V., SHARMA, R., AND HUANG, T. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *PAMI 19,* 7 (July), 677–695.

PAVLOVIC, V. I., SHARMA, R., AND HUANG, T. S. 1996. Gestural interface to a visual computing environment for molecular biologsts. In *FG96* (Killington, Vt.), 30–35.

PRILLWITZ, S., LEVEN, R., ZIENERT H., HANKE, T., AND HENNING, J. 1989. *Hamburg Notation System for Sign Languages—An Introductory Guide*. Signum, Hamburg.

QUEK, F. 1995. Eyes in the interface. *Int. J. Image Vis. Comput. 13*, 6 (Aug.), 511–525.

QUEK, F. 1996. Unencumbered gestural interaction. *IEEE Multimedia 4*, 3, 36–47.

QUEK, F. AND BRYLL, R. 1998. Vector coherence mapping: A parallelizable approach to image flow computation. In *ACCV*, Vol. 2, Hong Kong, 591–598.

QUEK, F. AND MCNEILL, D. 2000. A multimedia system for temporally situated perceptual psycholinguistic analysis. In *Measuring Behavior 2000*, Nijmegen, NL, 257.

QUEK, F., BRYLL, R., ARSLAN, H., KIRBAS, C., AND MCNEILL, D. 2001. A multimedia database system for temporally situated perceptual psycholinguistic analysis. *Multimedia Tools Apps.* in Press.

QUEK, F., MA, X., AND BRYLL, R. 1999. A paralle algorithm for dynamic gesture tracking. In *Proceedings of the ICCV'99 Workshop on RATFG-RTS* (Corfu), 119–126.

QUEK, F., YARGER, R., HACHIAHMETOGLU, Y., OHYA, J., SHINJIRO, K., NAKATSU., AND MCNEILL, D. 2000. Bunshin: A believable avatar surrogate for both scripted and on-the-fly pen-based control in a presentation environment. In *Emerging Technologies, SIGGRAPH 2000* (New Orleans) 187 (abstract) and CD–ROM (full paper).

SCHLENZIG, J., HUNTER, E., AND JAIN, R. 1994. Recursive identification of gesture inputs using hidden Markov models. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision* (Pacific Grove, Calif.).

SOWA, T. AND WACHSMUTH, I. 1999. Understanding coverbal dimensional gestures in a virtual design environment. In *Proceedings of the ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, P. Dalsgaard, C.-H. Lee, P. Heisterkamp, and R. Cole, Eds., Kloster Irsee, Germany, 117–120.

SOWA, T. AND WACHSMUTH, I. 2000. Coverbal iconic gestures for object descriptions in virtual environments: An empirical study. In *Post-Proceedings of the Conference of Gestures: Meaning and Use* (Porto, Portugal).

TRIESCH, J. AND VON DER MALSBURG, C. 1996. Robust classification of hand postures against complex backgrounds. In *FG96* (Killington, Vt.), 170–175.

WEXELBLAT, A. 1995. An approach to natural gesture in virtual environments. *ACM Trans. Comput. Hum. Interact. 2*, 3 (Sept.), 179–200.

WEXELBLAT, A. 1997. Research challenges in gesture: Open issues and unsolved problems. In *Proceedings of the International Gesture Workshop: Gesture & Sign Language in HCI*, I. Wachsmuth and M. Frohlich, Eds., Springer, Bielefeld, Germany, 1–11.

WILSON, A. D., BOBICK, A. F., AND CASSELL, J. 1996. Recovering temporal structure of natural gesture. In *FG96* (Killington, Vt.), 66–71.

YAMATO, J., OHYA, J., AND ISHII, K. 1992. Recognizing human action in time-sequential images using hidden Markov model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 379–385.

Received January 2001; revised December 2001; accepted January 2002