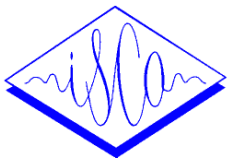

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 173-178.

Gesture marking of disfluencies in spontaneous speech

Yelena Yasinnik, Stefanie Shattuck-Hufnagel* & Nanette Veilleux***

* Massachusetts Institute of Technology, Cambridge, MA USA
** Simmons College, Boston, MA, USA

Abstract

Speakers effectively use both visual and acoustic cues to convey information in speech. While earlier research has concentrated on the association of visual cues (provided by gestures) with fluent prosodic structure, this study looks at the relationship between visual cues, prosodic markers and spoken disfluencies. Preliminary results suggested that speakers preferentially perform gestures in the eye region in spoken disfluencies, but a more careful frame-by-frame analysis capturing all gestures revealed that movements of the eye region (blinks, frowns, eyebrow raises and changes in direction of eyegaze) occur with high frequency in both fluent and non-fluent speech. The paper describes a method for frame-by-frame labelling of speech- accompanying gestures for a speech sample, whose output can then be combined with independently derived labels of the prosody. Initial analysis of 3 minute samples from two speakers reveals that one speaker produces eye movements in association with disfluencies and the other does not, and that this tendency does not result from alignment of brow gestures with pitch accents.

1. Introduction

This work focuses on the interaction among a) spoken disfluencies, b) visual cues provided by upper face movements, and c) prosodic prominence i.e. the perceptual salience provided by intonationally-cued phrase-level prominences called pitch accents. This relationship is of interest because it will tell us something about the underlying speech production planning process (i.e. about how speakers signal to listeners that a disfluency has occurred) and also because it may be of practical use in devising recognition systems that take advantage of such cues, as well as in developing on-screen personas that behave in a natural-looking ways. It is also possible that different kinds of disfluencies are associated with different kinds of gestural markers; such a finding would be particularly useful in automatic recognition and would have interesting implications for planning behavior. This study describes a labelling method for obtaining fine-grained information about the temporal relations among these three kinds of phenomena, and describes several preliminary results that illuminate the gesture-disfluency relationship.

2. Background

Initial studies of the relevant cues to prosodic prominence and phrasing in speech focused largely on acoustic cues such as f0, duration and amplitude. More recently, research on visual cues to prosodic events has contributed to our understanding of the complexity of human speech understanding and speech production. For example, visual cues from a speaker's gestures have been found to accompany the acoustic markers of prosodic events in speech, such as the phrase-level prominences called pitch accents. In particular, a number of investigators have reported an association between eyebrow raising and pitch accents (Keating et al. [5]). Yasinnik et al. [11] found that non-facial gestures produced by other non-speech articulators, such as the hand or head, are also associated with prosodic prominence; their study examined a certain type of gesture, which they define as a "hit" (i.e. a movement with an abrupt end point). Cavé et al. [2] reported an association between eyebrow movement and F0 rises, and showed that the two cues do not universally co-occur. Their results suggest "that eyebrow movements and fundamental frequency changes are not automatically linked (i.e., they are not the result of muscular synergy), but are more a consequence of linguistic and communicational choices".

One such communicational choice, possibly cued by both acoustic and visual means, is prosodic prominence. House et al. [4] found that the perception of prominence could be influenced by accompanying visual cues. In twelve acoustically identical stimuli, the presence/absence of head-nods and eyebrow movements of a computer-generated talking head influenced listeners' perceptions of which word was most salient. Although head-nods seemed to influence listeners more than eyebrow movements, eyebrow movements also significantly enhanced the perceived prominence of the closest pitch accented word. Similarly, Krahmer et al, [8] found that eyebrow movements did indeed contribute to determining which of two adjacent words was prominent, although in this case the presence/absence of an acoustic marker (pitch accent) was a more effective cue. Granström [3] found visual (again, facial) cues can indicate to human listeners whether an automatic system (as a travel agent talking head) has correctly understood the speaker. Taken together, these findings raise the possibility that spoken disfluencies may be marked gesturally, to help the listener identify the disruption.

3. Method

3.1. Database

Samples were excised from commercially available DVD recordings of 4 academic lecturers. Professional lectures were selected because they provide a ready source of large amounts of semi-spontaneous speech produced by speakers who are practiced communicators and recorded with high-quality audio. The main disadvantage of this corpus for our purposes is the fact that in some regions the speaker's body is either turned aside or not entirely visible because the producer chose to use a close-up frame or because other graphic material for the course is displayed instead. This disadvantage is easily overcome, however, because of the availability of many hours of recordings from each speaker. All 4 lecturers were male and appeared to be speakers of American English (M1am, M2am, M3am, M4am). The samples were approximately 40 minutes (M1), 9 minutes (M2), 11 minutes (M3), and 7.5 minutes (M4). Corresponding video and sound samples were transferred to a MacIntosh computer for gesture and speech labelling.

3.2. Preliminary Observation to Determine Relevant Gestures

The initial phase of the research involved watching the video files of all four speakers using iMovie and iDVD, locating speech errors and documenting any gestures – head, face, hands, and body - that occurred in the region containing the error, any editing remarks and the correction if one was made. No precise speech-gesture alignment was done at this point, but we noted the non-speech gestures that occurred in conjunction with each error.

The gestures we observed were sorted into categories according to their type, e.g. blink, shake, etc. Six main categories emerged:

- **Eyebrow raise**
- **Frown**
- **Blink**
- **Eyegaze transfer**
- **Shake (of the hand or head)**
- **Freeze –** an abrupt stop in the train of gesturing that was not preceded by relaxation (see below for discussion of gesture segment types, such as preparation, stroke, hold and relaxation).

Two aspects of this initial set of observations about disfluency-associated gestures are notable. First, many of the gestures involve the eye region: brow raises, frowns, blinks and eyegaze transfers. Two-thirds of the gestures that occurred in 59 disfluent regions (49/72) involved the eyes. Second, there was a striking absence of a type of gesture which occurs commonly in fluent speech, i.e. movements of the head or hands characterized by short sharp end points, that we have designated as 'hits' (Yasinnik et al. [11]). The movements of the head or hands that were observed in disfluent regions using this informal method were not the usual single-movement 'hits'; instead, we observed shakes (i.e. repeated short movements back and forth) or gestures that were temporarily frozen. It is possible that the gestures that were 'frozen' in disfluent regions would have been hits if completed, but in error regions they were not completed. Based on these preliminary observations, we formed two hypotheses about the relationship between disfluencies and speech-accompanying gestures in these four male speakers of American English: 1) disfluent regions in the speech are marked by gestures involving the eye region, and 2) they are not marked by hits. The remainder of this paper is focused on the first hypothesis.

These initial observations raised questions which require a more fine-grained method of analysis, one that permits investigation of the detailed alignments among disfluencies, speech-accompanying gestures and prosodic elements such as pitch accents. Two questions in particular arise about how to interpret the findings from the coarse-grained analysis. First, is the predominance of gestures that involve the eye region in disfluent regions unusual, or do eye movements occur freely and often throughout the speech samples? Second, is this predominance due to the fact that eye movements occur on pitch-accented syllables and thus error corrections (which are likely to include contrastively pitch accented syllables) are also likely to be associated with eye movements?

The labelling method we adopted for this more fine-grained analysis was developed in earlier studies of the alignment of gestures with aspects of prosodic structure such as prominences (pitch accents) and constituent boundaries (e.g. intonational phrase boundaries). In this method, the video file and audio file for the sampled lecture are separated, and labelled independently (by different labelers). The sound file is labelled for word alignments, prosody (e.g. intonational phrases and pitch accents) and disfluencies, and the video file for frame-by-frame gestural events, capturing the various subsections of a gesture, such as the preparation stage, the gesture itself (with optional hold), the relaxation stage and optional pause before the next gesture (McNeill, [9]). Aligning these two sets of independently transcribed labels provides a way of testing for co-occurrence of the gestures with the speech at the syllable-by-syllable level, and this analysis can be carried out with confidence that the results are not determined by any perceptual bias e.g. toward aligning gestural strokes with auditorily prominent syllables, etc.

3.3. Labelling

Because this aspect of the work is very time consuming, we selected shorter three-minute samples from two of the original speakers, M2am and M4am, for fine-grained labelling. These samples contained a substantial number of disfluencies, 16 for M2 and 16 for M4, and minimal video disruption from lecture-related graphics. For each sample, the sound file was transcribed and the words aligned with the wave form and spectrogram using xwaves; this representation formed the basis for prosody labelling of pitch accents and boundary related tones using the ToBI system, and the disfluent regions. The video file was labelled frame by frame for the onset of each subsection of each gesture. The labelling method for both files is described in some detail here, because it gives a flavor of the level of precision that is captured by these labels, and provides a clear picture of some of the challenges that can arise with labelling speech and gestural phenomena.

Gesture labelling. Gestures were labelled in the video file, without listening to the sound, using Anvil 4.5.2 [7]. This software allows the creation of multiple time- and video-aligned tiers for labelling gestures performed by different articulators: hands, head, eyes, and eyebrows. The tiers were displayed with tick marks for every second of the video and smaller tick marks equivalent to one video frame - $1/30^{th}$ part of a second. Every frame corresponded to one image in the video.

Within each tier, the labeller marked a beginning and an end of a region and designated it for a certain gesture with a label and an optional comment. The criteria for marking the onset and offset of a gesture were adopted from criteria for onsets and offsets of hand gestures described in Yasinnik et al. [11]:

- Onsets were marked at the frame where shape of the articulator began to change (e.g. narrowing of eyes during a blink), or the articulator's position began to change (e.g. head turn during a head shake), or the articulator's location began to change (e.g. hand moving from one place to another) which often was accompanied by blurring of the image in the video frame. The blurring provided a useful clue to the location of the onset video frame.
- Offsets were marked at the frame where the articulator stopped moving (this generally corresponded to a clearer image than in the surrounding frames), or at the frame just before the next change in articulator shape or position, which usually signaled the relaxation stage, but sometimes was simply an onset of the next gesture.

The labels in all tiers were saved together as an Anvil annotation file, which provided a text summary of time-stamps for onsets and offsets of gestures, as well as the labels for all marked gestures.

Reliability. Two labellers separately labelled the hand gestures in a 30-second segment of one of the video files. For each gesture, several phases were identified: preparation, stroke (sometimes followed by hold) and relaxation (for related discussion see McNeill [9] and Kendon [6]).

Out of 85 and 84 labels provided by Labeller 1 and Labeller 2 respectively, 71 markings (84%) agreed on a label and a time-stamp within one frame and 75 markings agreed within three frames. Inter-labeller disagreements arose largely from two sources: a) Labeller 2, who was more familiar with this speaker's gestures from earlier labelling experience, understood that during a pause in gesturing, this speaker often tapped his hands together while in neutral position, and labelled this region as a pause, while the Labeller 1 labelled each tap as a separate gesture event, and b) during one diagonal upward hand movement (accompanied by shaking), Labeller 2 broke the gesture into two shakes in different spatial locations, separated by a preparation stage. The limited nature of these disagreements suggests that more specific definitions of gestural onsets and offsets will increase this already high level of agreement.

Prosody labelling. The prosody of the spoken utterances was independently labelled from the sound files, by an experienced labeler using the ToBI system (http://www.ling.ohio-state.edu/~tobi/ame_tobi) from sound files displayed in Praat (see Boersma [1]). Aspects of the labels used here include pitch accent locations (i.e. the words and syllables marked with intonational phrase-level prominence) and intonational phrase boundaries. Experienced ToBI labellers transcribed the words (including the symbol PAU for each perceptually-noticeable silence) in the Praat TextTier file "words", using the waveform and spectogram to align the words with the sound. They also transcribed the pitch accents and boundary-related tones of the utterances in corresponding "tones" and "breaks" TextTier files, while listening to the sound files and viewing time-aligned f0 tracks, which had been created with the Praat pitch-tracking algorithm. The "words", "breaks", and "tones" TextTier files provide a time-stamp for the beginning and ending of each transcribed word, part of a word – if the word is cut off, and of each pause, as well as a time-stamp and label for each marked tonal element and intonation phrase boundary.

Disfluency labelling. In the miscellaneous TextTier, disfluent regions were assigned one of the four categories summarized below; for a related disfluency classification scheme see Shriberg [10].

- **Filled pause** – contains filler words *um* or *uh*, (e.g. *about that uh equation*);
- **Bobble** - contains repetition of a word or part of a word, (e.g. *many k- complications, forms the- the basis*); this category might be called a stutter, but this term has wide use in the literature for a type of speech pathology, and in addition, the two renditions of the target were sometimes separated by a filled pause, e.g. *mol-uh-molecule*, which is not the intuitive sense of what the term 'stutter' means;
- **Substitution** - contains substitution of a linguistic element often followed by a correction, i.e. an incorrect word, syllable, or sound, often drawn from nearby context (e.g. *har- artists had, ungovernmentally sh- sanctioned*)
- **Other disfluencies** such as a change of plan after a phrase has started, (e.g. *They, in fact, have all- …, they actually had to join the group*) or one whose type is ambiguous (e.g. *the b- tennis ball*, where the word fragment *b-* can reflect either a change of plans, i.e. a plan to say *the ball* replaced by a plan to say *the tennis ball*, or a linguistic substitution like *the bennis tall*, that was interrupted and corrected by the speaker before it was fully pronounced).

The 'miscellaneous' TextTier file provides a time-stamp for the ending of the most perceptually disfluent word, or part of a word, within each error. Disfluencies could be divided into the standard subsections of reparandum (region requiring repair), editing and repair, although defining the extent of the repair in cases of a change in plans is not straightforward.

3.4. Alignment

This set of labels provides a complete record of the gestures made by the speaker during the sample, and also reveals the precise alignment of gestures, prosodic prominences and disfluencies. Preliminary analysis of these data for the two 3-minute samples focused on two questions: (1) is the preponderance of eye-region gestures in disfluent regions a characteristic of disfluencies, or does it arise because eye-region gestures occur with high frequency throughout both fluent and disfluent speech? (2) do eye-region gestures of a particular kind, eyebrow hits, align differently with the prosodic elements of spoken utterances than other kinds of speech-accompanying gestures, and (3) do eyebrow hits occur preferentially in association with disfluencies?

4. Results

Results from the broad marking of gestural events in disfluent regions for the four speakers were described above. We noted the predominance of eye-region gestures (eyebrow raise, frown, blink, or eyegaze transfer), and absence of hand or head 'hit' gestures except for occasional shaking; any ongoing head or hand movements that might have been hits froze during the error interval. In addition, only 10 of the 59 errors (17%) were not marked by any gesture, and most of these non-gesture-marked errors were Filled Pauses (7 / 10). We note that these errors are by definition marked in the acoustic signal (in the form of *um* or *uh*), which is consistent with the hypothesis that eye-region gestures during disfluent regions are less likely in circumstances where the speech signal does not provide unambiguous cues to the occurrence of a disfluency, or does not provide them early enough in the disfluent event to prevent initial confusion on the part of the listener.

(1) How prevalent are eye gestures for these speakers? Analysis of the fine-grained gesture labels for three minutes of speech for each of two speakers (M2am and M4am) confirmed the general prevalence of eye movements in disfluent regions obtained from informal observation: 70+% of errors co-occurred with an eye gesture of some kind. 70% (7/10) for M2am, and 73% (22/27) for M4am. However, these results must be compared to the number of eye gestures accompanying equivalent non-disfluent speech. Since there is some difficulty in determining precisely what an appropriate comparison interval in fluent speech would be, we used a simple 3-word interval as an approximation since errors generally involve about three words. This provides a rough impression of how widespread eye gestures are throughout the samples. In the M2am sample, there were 313 eye gestures distributed over about 820 fluent words or an average of 1.14 eye gestures for every three word interval. Similarly, there were 188 eye gestures distributed over about 500 fluent words in the M4am sample or an average of 1.13 eye gestures for every three word interval. This indicates that eye gestures are frequent, commonly occurring in both fluent and disfluent speech.

This finding suggests that eye gestures as a class may be too broad a category to gainfully map to errors. Results from the coarser labels suggested that one type of eye movement, brow raising, is particularly common in disfluent regions, and Keating [5] has reported that brow movements can be associated with pitch accents. As a preliminary to testing the hypothesis that brow movements occur preferentially in conjunction with disfluencies vs. with pitch accents, we examined the distribution of gestural 'hits' by three different articulators: hands, head and eyebrows.

(2) Do gestural 'hits' of eyebrows, head and hands align with pitch-accented syllables? If some eye movements occur disproportionately often in disfluencies, it may be because they tend to occur on pitch-accented syllables, and errors (or at least corrections) are likely to contain contrastively accented syllables. If this is the case, then we expect to find that eyebrow hits occur more reliably on accented syllables than do hits produced by other articulators such as the head or hands. To test this hypothesis, we examined the 30 millisecond time segment of each hit frame in the sound waveform of the lecture. Many aspects of the alignments are of interest, including the hit frame's location in relation to the corresponding word, pitch accents, location of pitch accents in the syllabic structure of the word, and the phrasal structure. Here we focus on whether the 30-ms frame of each gestural hit overlapped with the duration of a pitch-accented syllable.

This alignment can take one of four forms: (1) *Early PAcc* if the hit frame occurs in the syllable preceding the speech prominence; (2) *Yes* if the frame occurs in the same syllable as the prominence; (3) *Foot* if the hit frame occurs in the syllable following the prominence and if that syllable is weak; and (4) No if the hit frame location has no relation to any syllable marked with a pitch accent prominence. An additional comment of *Maybe* was added to each label when the corresponding pitch accent label was marked as uncertain (i.e. *? In the ToBI system), indicating that the labeler was unsure whether there was a pitch accent. NB: although the foot is often defined as a strong syllable plus up to two weak following syllables within a word, e.g. *WOman*, we adopted a unit more like the Abercrombian foot, for which the weak syllable can be part of the following word (e.g. *send him* in *send him out*, *bake a* in *bake a pie*, or even *wrote a-* in *wrote about*).

The timing of hits and pitch accents were strongly correlated for this sampled speaker. The very high Pearson's Index, 0.98, established that gesture hits and pitch accents correlate positively with greater than 99% confidence ($p < 0.01$). Overall, 53% of hits occurred in pitch accented syllables, 24% occurred in the second syllable of the foot, and 13% occurred in other non-prominent parts of the speech signal. This data is consistent with other previously sampled speakers, but a closer examination of the alignment results revealed some interesting differences between gestural articulators.

While the correlation for hand hits with pitch accents was high (59% aligned with the accented syllable and an additional 19% with the second syllable of the foot, **Figure 2**) and head hits also showed a high correspondence (44% on the accented syllable and 42% on the following weak syllable, **Figure 1**), the eyebrow hits produced very different results: hits aligned equally with the accented syllable, the following weak syllable and with no accent-related syllable (**Figure 3**). This difference suggests that the relatively tight relationship between pitch accents and gestural hits observed for head and hand gestures is not observed for eyebrow gestures, at least for this speaker's sample. This supports the hypothesis that when brow gestures occur in disfluent regions, it is not just because they tend to be associated with pitch accents.

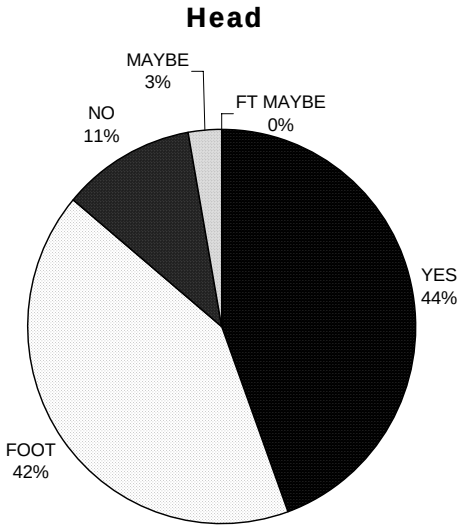

Figure 1: Percentage of head hit frames aligned with Pitch Accents (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? Pitch accent (Maybe) and with No pitch accent.

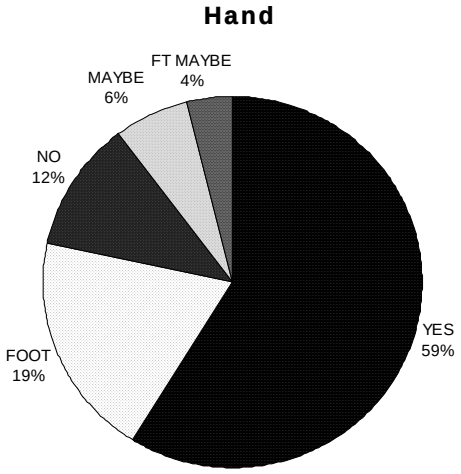

Figure 2: Percentage of hand hit frames aligned with Pitch Accents (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? Pitch accent (Maybe) and with No pitch accent.

We note also that the percentage of head hits aligned with the second syllable of the foot is large compared the percentage for hand hits. It is possible that this reflects the greater inertia of the relatively massive head, which may take longer to move into position.

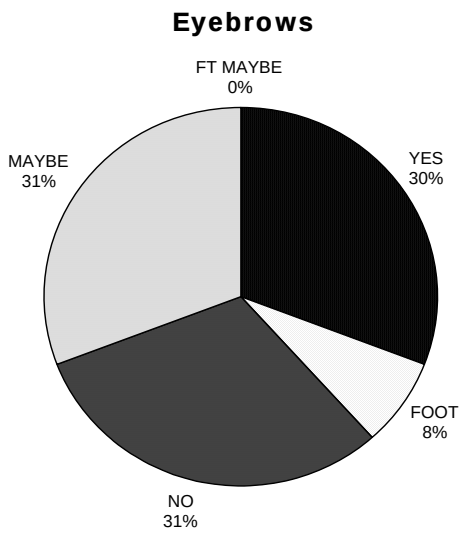

Figure 3: Percentage of eyebrow hit frames aligned with a Pitch Accent (Yes), with the Foot (an unstressed syllable after the pitch accent), with a *? pitch accent (Maybe) and with No pitch accent.

(3) Do eyebrow hits occur preferentially with disfluencies? Because eyebrow hits are distributed differently from hand and head hits in general in these two speech samples, we also analyzed their alignment with disfluent regions. To do this we determined the location of the 30 millisecond frame of each eyebrow hit in the speech waveform, and examined the nearby speech to see if a disfluency occurred. This analysis was confined to large, salient eyebrow hits; some eyebrow movements were rather small and difficult to distinguish from changes in light glinting off the spectacles of the speakers, and it was important to be sure we were looking at the distribution of eyebrow hits that would be salient to the watching listener. Results were strikingly different for the two speakers. Speaker M2am showed a tendency for eyebrow hits to occur in disfluent regions: of his 17 salient eyebrow hits, 11 occurred in conjunction with a disfluency and 6 did not. Speaker M4am, however, showed a very different pattern: of his 18 salient brow gestures, only 2 occurred near a disfluency and 16 did not. Since the number of disfluencies and the number of salient brow hits were similar for the two speakers (16 and 16 errors, 17 and 18 brow hits), these observations suggest that these two speakers are using eyebrow movements in different ways.

5. Discussion

The results described above confirm findings from earlier studies showing that speech-accompanying gestures are not randomly distributed in the speech signal, but in many cases are systematically aligned with speech events. In particular, the alignment of head and hand hits with pitch accents for speaker M2am is consistent with earlier work. However, this speaker does not align his brow hits with pitch accented syllables as reliably as his head and hand hits; instead, he tends to align them with disfluent regions. Thus, results for this speaker support the hypothesis that at least some types of eye-related gestures occur preferentially with disfluencies. However, the finding that speaker M4am does not align his brow hits with disfluent regions highlights the importance of surveying multiple speakers to determine which aspect of the gesture-disfluency relationship are general across speakers and which are idiosyncratic. It is possible that gesturing patterns are particularly variable across individuals, more so for example than some aspects of prosodic and syntactic usage.

A top priority for future work concerns the question of whether different types of disfluencies are marked by different kinds of gestures. The hypothesis that disfluencies that do not contain immediately obvious cues to their disfluent nature, such as substitutions, are preferentially marked by eye-region movements (because listeners will be focused more on the eye region of the talker) was not confirmed, since eye gestures occur freely throughout the speech samples we examined. However, it is still possible that errors that might initially mislead the listener into thinking they are part of fluent speech, such as word substitutions, are marked by certain types of eye gestures, while errors that are acoustically distinct (such as the repeated words/sounds of a bobble or the filler items of a filled pause) may be less consistently eye-gesture marked.

Another line of investigation will be to determine the extent to which all four speakers align their brow movements with pitch accents; initial observation suggests that Speaker M4am, whose brow hits were not associated with disfluencies, tended to align them with pitch accented syllables instead. If this initial impression is confirmed, it will reinforce the necessity of studying differences as well as similarities in the gestural marking of disfluencies by individual speakers. In addition, some of M2am and M4am brow movements were not discrete hits but raises that remained up for an extended interval. These gestures may be related to discourse structure.

Finally, we plan to pursue further the informal observation that there is a paucity of single hits in disfluent regions although these gestures are common in the fluent speech of these speakers.

6. Acknowledgements

Work supported by NIH/NIDCD grant RO1 DC00075, the Keith North Fund at the Speech Group at MIT/RLE, and MIT's Undergraduate Research Opportunities Program. Technical assistance and support from Helen Hanson and Michael Kipp are gratefully acknowledged, along with the time and effort invested in the project by Margaret Renick, Jessie Wang and Alicia Patterson.

7. References

[1] Boersma, P. 2001. PRAAT: A System for Doing Phonetics by Computer. *Glot International*, vol. 5, pp. 341-345.

[2] Cavé, Christian, Isabelle Guaïtella, Roxane Bertrand, Serge Santi, Françoise Harlay, Robert Espesser. About the Relationship Between Eyebrow Movements and F0 Variations. 1996. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, vol. 4, pp. 2175-2179.

[3] Granström B, House D. & M. G. Swerts. 2002. Multimodal Feedback Cues in Human-Machine Interactions. In B. Bel & I. Marlien (eds.), *Proc. Speech Prosody 2002 Conference*. Aix-en-Provence, France, pp. 347-350.

[4] House, David, Jonas Beskow & Bjorn Gransrom. 2001. Interaction of Visual Cues for Prominence. *Working Papers 49*, Dept. of Linguistics, Lund University, Sweden, pp. 62-65.

[5] Keating, P., M. Baroni, S. Mattys, R. Scarborough, A. Alwan, E. Auer, & L. Bernstein. 2003. Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English. *Proc. 15th International Congress of Phonetic Sciences*, 3-9 August 2003, Barcelona, Spain, pp. 2071-2074.

[6] 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*, The Hague: Mouton. pp. 207-227.

[7] Kipp, Michael. 2001. Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 1367-1370.

[8] Krahmer, Emiel, Zsofia Ruttkay, Marc Swerts & Wieger Wesselink. 2002. Pitch, Eyebrows and the Perception of Focus. *Proc. Speech Prosody 2002*, 11-13 April 2002, Aix en Province, France.

[9] McNeill, D. 1992. *Hand and Mind: What Gestures Reveal About Thought*. Univ. Chi.: Chicago.

[10] Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, USA, vol. 3.

[11] Yasinnik, Yelena, Margaret Renwick & Stefanie Shattuck-Hufnagel. 2004. The Timing of Speech-Accompanying Gestures with Respect to Prosody. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 11-13 June 2004, Cambridge, MA.