Head movements and prosody in multimodal feedback

Max Boholm
SCCIIL (SSKKII)
Department of Applied Information Technology
University of Gothenburg
Gothenburg, Sweden
max.boholm@gu.se

Gustaf Lindblad
SCCIIL (SSKKII)
Department of Applied Information Technology
University of Gothenburg
Gothenburg, Sweden
gustaf.lindblad@sskkii.gu.se

Abstract

The study analyses the relation between words, including their prosodic features, and head movements in communicative feedback, i.e. unobtrusive vocal and gestural expressions which convey information about ability and willingness to continue, perceive, and understand, as well as attitudes and emotions. Examples are words such as *m* and *okay*, and head movements such as nods and shakes. Six recorded first acquaintance conversations in Swedish have been analyzed. Initial direction, repetition, start time, and duration of head movements has been identified by frame-by-frame video analysis. Start time, duration, F0-contour, and pitch of vocal-verbal feedback were analyzed. Main results of the study are: first, multimodal nods more frequently start before or at the same time as words, than words starting before nods. Second, nods have longer duration when produced with words than without. Third, certain words are typically associated with certain nod types, e.g. *okay* with up nods, and *m* with repeated nods. Finally, certain prosodic patterns are more associated with certain nod types, e.g. rising pitch and longer durations with single up nods, and falling or flat pitch with repeated down nods.

1 Introduction

It has often been recognised that gestures can serve to express many of the functions that are known to be expressed by prosody. For example, emphasis of words and phrases in speech can be achieved by both prosodic features and so called "batonic" gestures with hand or head (Bull and Connelly, 1985; Kendon, 2004; McNeill, 1985). Today a growing literature suggests a tight connection between prosodic features of speech and the gestures that accompany speech. This multimodal interplay is sometimes discussed under the terms of *optical phonetics* (Scarborough et al 2009) or *visual* (or *audiovisual*) *prosody* (Graf et al, 2002; Krahmer and Swertz, 2009; Munhall et al, 2004; Swertz and Krahmer, 2010). Words that are made prominent by acoustic means are often accompanied by head and eyebrow movements (Swertz and Krahmer, 2010). Graf et al (2002) report that pitch accents are strongly correlated with accompaniment of head movements. Scarborough et al (2009) found that facial movements were larger and faster with stressed words. Related to these findings that strongly suggest a co-activation of acoustic and gestural means in producing prominence of a linguistic component, Cavé et al (1996) observed a kind of audio-visual isomorphism. They found that the F0 rises were accompanied by raised eyebrows in 71% of the cases (Cavé et al, 1996). It has also been demonstrated that gestural visual cues play an important role for the perception of a word as prominent, and even that gestural accompaniment facilitate speech perception, comprehension and intelligibility as well as the experienced naturalness of Embodied Communicative Agents (see Munhall et al, 2004; Granström and House, 2005; Moubayed et al, 2010). Taking into account the interaction of lexical and prosodic features, as well as timing, have been shown to improve recognition of feedback head movements in human-computer interfaces (Morency et al 2007).

The present study analyses the relation between words, including prosody, and head movements in communicative feedback. Feedback is defined as unobtrusive vocal and gestural expressions used in communication to inform an interlocutor about the ability and willingness to (i) continue the interaction, to (ii) perceive, and

(iii) understand what is communicated, and (iv) in other ways attitudinally and emotionally react (see Allwood, 1988; Allwood et al, 1992; Allwood et al, 2007). Types of vocal-verbal feedback in Swedish include feedback words, feedback phrases, feedback clauses such as *jag förstår* ('I understand') and repetition of what the interlocutor just said (other-repetition). Feedback words, in turn fall in two sub-types (see Allwood 1988): primary and secondary feedback words. Primary feedback words are words which are used to express feedback, i.e. the basic communicative functions (i-iv) above, but which cannot be used as predicates, attributes or adverbs. Examples of primary feedback words in Swedish are *m* ('m'), *ja* ('yes'), *nä* ('no'), *jo* (contrastive 'yes') and *okej* ('okay'). Secondary feedback words are words which in addition to function as feedback can be used as predicates, attributes or adverbs (and are tentatively more commonly used with such functions). Examples of secondary feedback words in Swedish are adjectives and adverbs such as *precis* ('precisely'), *bra* ('good') and *exact* ('exactly'). These types of vocal-verbal feedback can be used alone or in combination, forming units, which in turn can have different positions in an utterance: (i) single position, i.e. constitute the entire utterance, (ii) initial position, (iii) medial position, and (iv) final position. Of these, single and initial positions are characteristic and the most common for feedback.

Examples of gestural feedback are head nods, head shakes, smiles, raised or frowning eyebrows and shoulder shrugs. Using vocal-verbal and gestural feedback in combination results in multimodal feedback, e.g. a feedback word *ja* ('yes') in combination with a nod.

The following research questions are addressed: What are the timing relations between head movements and words in multimodal feedback contributions, i.e. do head movements start before, at the same time or after words, or vice versa? What prosodic features (F0-curve, duration, pitch) of vocal-verbal feedback are found when produced with versus without accompanying head movements, and more specifically, in relation to different kinds of head movements, i.e. in terms of their initial direction (e.g. up, down), repetition (repeated, single) and duration? Which feedback words co-occur with which prosodic patterns and with which types of head movements?

2 Method

Data for analysis consist of six video and audio recordings of dyads in (spontaneous and natural) first acquaintance conversations. Audio data was recorded with individual microphones for each speaker to facilitate acoustic/prosodic analysis. Video data was recorded using a three camera set up, with one camera taking in the whole scene, and two cameras focusing on the head and torso of each speaker respectively. The subjects had never met prior to the recording session, and were instructed to get to know each other during approximately 8 minutes. All subjects except one were university students. Of the six recordings, four are male-male and two are female-female conversation. Two of the speakers take part in two conversations each (with different partners), so the empirical material comprises 10 different speakers (six males and four females), in total.

Head movements have been coded by manual frame-by-frame analysis of the video recordings, identifying the following features: (i) type of head movement: head nod (vertical movement of the head, where the chin's distance to the torso varies as the head goes up and down), head shake (horizontal movements of the head, turning the head from side to side), head tilt (vertical and horizontal movements, tilting the head from side to side) or other; (ii) initial direction (in the case of head nods): up or down; (iii) repeated or single movement; and (iv) start time, end time and duration. Each frame of the recording is 40 milliseconds (ms), hence measures of time for head nods are measured with a level of detail of 40 ms.

Vocal-verbal feedback were analyzed using Praat and Audacity, identifying starting point, duration, general shape of the F0-curve and mean frequency of F0. Measurements of the average pitch of the highest and lowest 30 ms portions of the F0 curve were also taken for every utterance.

All analyzable cases of vocal-verbal feedback were categorized as one of eight different F0-curve types. The types were: complex, complex-down, complex-up, down, down-up, flat, up, and up-down. Up and down are to be interpreted as rising and falling pitch respectively. The categorization of vocal-verbal feedback was straightforward for most cases, as the shapes of the F0-curves clearly fell into one category or another. A statistical relation was used to decide if a curve was flat: any curve where the difference between the highest and the lowest 30 ms portions was less than 5% was deemed to be flat,

since such a small difference in pitch would not be audibly noticeable. The complex-up and complex-down categories were used for curves that had a general rising or falling shape, but with some irregularities of one kind or other. The complex category was used for curves that were judged not to fit into any of the other categories (22 out of 618 analyzed instances).

Certain values where derived from these measurements. An average pitch value of the F0 was calculated for every speaker (*Speaker Pitch*), based on the average of all of that speaker's vocal-verbal feedback. Subsequently, the average pitch of every vocal-verbal feedback unit was compared to the Speaker Pitch to describe its relative pitch (*Frequency Deviation*). For vocal-verbal feedback with a non-flat F0-curve, the difference between the highest and the lowest average was calculated as a percentage value (*Frequency Difference*).

108 out of 703 cases of vocal-verbal feedback were not analyzable in all prosodic dimensions, and 23 of these were not analyzable for any prosodic qualities at all. The most common reason for a unit not being analyzable is that sound from the other speaker is bleeding in to the signal, thus masking it. Instances of unanalyzed or partly analyzed vocal-verbal feedback were still used in cases where the affected prosodic dimensions were not of interest.

All pitch data should be fairly accurate within ±1 Hz. Duration data should be accurate within ±10 ms. The margin of error for comparative timing data is about ±40 ms or ±1 frame. Because the audio was recorded separately from the video, to ensure good audio quality, the two data streams had to be synchronized post recording. As the video is the master time track, the accuracy of timing relations is only as good as the time resolution of the video.

3 Results

3.1 General observations

Since the feedback system involves both vocal-verbal and gestural means, as well as different possible combinations of them, a variety of feedback types are possible. Based on the type of vocal-verbal component of the feedback contribution, if any, and the type of head movement, if any, the feedback types in Table 1 have been identified.

Head nods are by far the most common head movement used for feedback in the analyzed material, where 534 feedback head nods have been identified, while only 20 instances of other head movements with feedback function (e.g. shakes and tilts). Feedback head nods are more often co-produced with words (393 instances), than without words (141 instances): 74% vs. 26%. Inversely, vocal-verbal feedback is also more likely to be produced with feedback nods (393 instances, versus 290 instances produced without), but the difference is not as large: 58% vs. 42% (59% vs. 41% when including other head movements).

Vocal-verbal feedback (FB)	**Head movement**			**Tot.**
	Nod	Other	None	
Single primary FB word	297	15	227	539
Series of primary FB words	43	3	27	73
Combo of primary FB word(s) & secondary FB words (adverbs)	23	2	21	46
Combo of primary FB word(s) & other-repetition	4	0	9	13
FB clause	8	0	1	9
Other vocal-verbal (without FB word)	9	0	0	9
Single secondary FB word (adverb, adjective)	4	0	1	5
Primary FB word and OCM word	1	0	4	5
Other-repetition	2	0	0	2
Combo of secondary FB word (adverb) & other-repetition	1	0	0	1
Combo of primary FB word, secondary FB word (adverb) & other-repetition.	1	0	0	1
No vocal-verbal feedback (silence)	141	0	-	141
Total	534	20	290	844

Table 1. Number of instances of combinations of different kinds of vocal-verbal feedback and feedback head movements.

The two most common vocal-verbal forms of feedback are single primary feedback words, e.g. *m* ('m'), *ja* ('yes'), *nä* ('no'), *jo* (contrastive 'yes') and *okej* ('okay'), and primary feedback words used in series, e.g. *ja okej* ('yeah okay') and *ja ja* ('yeah yeah') (self-repetition). Furthermore primary feedback words are found in combination with secondary feedback words (e.g. adverbs), other-repetition and words for own communication management (OCM), e.g. *eh ja* ('um yeah'). Feedback that has a vocal-verbal component but lack a primary feedback

word altogether is uncommon. All these vocal-verbal types of feedback have initial positions of utterances, i.e. are followed by some non-feedback part, or they constitute the entire utterance themselves.

In addition to the contributions presented in Table 1, there are four cases of contributions in the recorded conversations which consist of a nod but where the vocal-verbal component is impossible to hear. None of these cases are considered for analysis below.

Due the meager data on other feedback head movement than head nods (only 20 instances in total), results on duration and timing focus only on nods (sect. 3.2 and 3.3). Also the comparison between head movements and prosodic features of speech will mainly, but not exclusively, focus on nods (sect. 3.5).

3.2 Timing of nods and words

In multimodal feedback contributions including nod and words (n=393), the nod can start before, after or at the same time as the word starts. That nod and word(s) start and/or end at exactly the same time is very unlikely, even though there are indeed two instances where this has been observed. This is of course subject to the level of detail of measurement, which in this study comes down to the marginal of a frame (40 ms). Consequently, in almost all cases the nod starts before the word(s) or the word starts before the nod, where the former being slightly more common than the latter (195 vs. 150 instances). However, since it is quite common that the nod and the word(s) start within a 120 ms time span,[1] i.e. almost at the same time, the following relations can be differentiated:

a) Nod starts more than 120 ms before word(s) (115 instances; 29% of the cases)
b) Nod and word(s) start within 120 ms span (146 instances; 37% of the cases)
c) Nod starts more than 120 ms after word(s) (instances 86; 22% of the cases)
d) In 46 cases the timing relation is unknown due to lack of reliable measurements (12% of all instances)

A majority of type c are produced with a gap (53 instances), i.e. the word both start and end before the nod starts. Less common, there are also 23 instances of gaps in the case of type a. This raises questions about the multimodality of such cases, and it is here argued that multimodality is a question of perceiver interpretation; that two communicative behaviors in different modalities belong together as a multimodal unit cannot be reduced to a simple question of co-occurrence in time.

3.3 Duration of nods

Feedback nods are longer when co-produced with words (multimodal), than when produced without words, see Table 2.

Nod type	With Words (MM)			Without words			MM nods are:
	M	n	St.d	M	n	St.d	
Repeated down nod	1201	153	724	940	84	469	28% longer
Single down nod	330	33	105	273	6	96	21% longer
Repeated up nod	1229	108	790	1028	40	420	20% longer
Single up nod	511	99	290	422	11	196	21% longer
Total	961	393	723	896	141	472	

Table 2. Mean duration in milliseconds (ms) and standard deviation of feedback nods in relation to co-production with words (multimodal, MM), or not.

Table 2 shows that for all nod types the nods which are produced with words are 20-28% longer in mean duration than those produced without words. As we shall see below, words do not have longer duration when they are accompanied by head movements, in general.

3.4 Head movement types and feedback words

A majority of the contributions under consideration here contain one or several primary feedback words, in different ways, e.g. as a single constituent of the vocal-verbal feedback, in series or together with secondary feedback words or other-repetition. (Exceptions are, for instance, contributions that as a vocal-verbal feedback part only consist of secondary feedback word or other-repetition, see Table 1).[2] Primary feedback words differ both in the extent that they do co-

[1] The 120 ms (three video frames) time span is chosen because it is larger than the error margin of the synchronization of video and audio, while still being an almost unnoticeable delay for a human observer.

[2] There are 684 contributions in the empirical material which contain at least one primary feedback word.

occur with head movements, and the types of head movements (and nod types) they do co-occur with.

First, considering the five most common feedback words, which all are primary, the prevalence of accompanying nods differ. The five most common feedback words in the material are: *ja* ('yes/yeah') (382 instances), *m* ('m') (148 instances), *okej* ('okay') (78 instances), *nä* ('no') (55 instances), and *jaha* ('yes, I see') (23 instances). The feedback word *ja* ('yes/yeah') is equally common together with head movement as without any. The word *nä* ('no') is slightly more common without head movements than with. The words *m* ('m'), *okej* ('okay') and *jaha* ('yes, I see') are more common together with head movement than without. See Table 3.

FB words	n	Without head movement	With head movement
ja	382	50%	50%
m	148	23%	77%
okej	78	24%	76%
nä	55	55%	45%
jaha	23	35%	65%

Table 3. The extent that the five most common feedback (FB) words *ja* ('yes/yeah'), *m* ('m'), *okej* ('okay'), *nä* ('no') and *jaha* ('yes, I see') are multimodal with respect to head movements.

Second, looking closer at the types of head movements that accompany these five words, further differences emerge, see Table 4.

FB words	Rep. down nod	Sing. down nod	Rep. up nod	Sing. up nod	Other head movem.
ja (n=191)	31%	10%	27%	29%	2%
m (n=114)	56%	10%	29%	4%	2%
okej (n=59)	15%	2%	34%	46%	3%
nä (n=25)	24%	0%	8%	40%	28%
jaha (n=15)	0%	0%	13%	80%	7%

Table 4. The relation between the five most common feedback (FB) words *ja* ('yes/yeah'), *m* ('m'), *okej* ('okay'), *nä* ('no') and *jaha* ('yes, I see') and different kinds of head movements in multimodal feedback.

The word *m* ('m') is strongly associated with repeated nods. For *m* ('m') co-production with repeated down nods and repeated up nods constitute 85% of its uses in contributions with head movement. Of the five words, *m* ('m') is the word which is strongest associated with repeated down nods. Both *okej* ('okay') and *jaha* ('yes, I see') are strongly associated with (single and repeated) nods which have an initial upward direction: 80% of *okej* ('okay') and 93% of *jaha* ('yes, I see'), which are produced with head movements, are produced with single or repeated up nods. Of the five words, *jaha* ('yes, I see') is the word which is strongest associated with single up nods (80%). The word *nä* ('no') is the only of the five words which is common together with other head movements than nods. The most common kind of head movement in question here is the head shake. It should however be noted that *nä* ('no') is more common with nods than with shakes. This results is to be interpreted in relation to the affirmative use of *nä* ('no') in response to utterances which contain negation (see e.g. Allwood et al 1992). When the word *ja* ('yes/yeah') is co-produced with head movements, it is overwhelmingly used with head nods, but lacks any strong association with a particular type of nod.

3.5 Head movements and prosody

This section discusses the prosodic features of word duration and pitch in relation to head movements in multimodal feedback. Above, feedback nods were found to have longer duration when accompanied by words, see section 3.3. Turning to the duration of vocal-verbal feedback, the trend that "multimodal is longer" does not seems to apply; cf. Allwood and Cerrato (2003) who found that feedback words were 20-40% longer when produced with head movements. Table 5 shows the mean duration of the five most common feedback words, in cases where these feedback words alone constitute the vocal-verbal feedback of a contribution, including all cases when this feedback is only a part of a contribution as well as constituting the whole contribution (see "Single primary FB word" of Table 1).

In cases of *m* ('m') and *okej* ('okay') as single feedback words, the difference in mean duration of them being co-produced with nod or not is minimal (only 1% difference in the case of *m* and 3% difference in the case of *okej*). The word *nä* ('no') as a single feedback word is slightly longer in duration when produced with head movement, than when it is produced without (9% longer), while *ja* ('yes') is slightly longer in duration when produced without head movement than with (15% longer). Of these single feedback words, only *jaha* ('yes, I see') is considerably longer when produced with head movement, than

without head movement (56% longer), but note that the instances of *jaha* ('yes, I see') are quite few.

Single FB word	n	With head movement			Without head movement		
		M	n	St.d	M	n	St.d
ja	273	217	132	75	250	141	137
m	120	225	93	68	222	27	57
okej	47	305	40	145	313	7	114
nä	32	255	13	121	233	19	94
jaha	11	397	7	122	255	4	55

Table 5. Mean duration of the five most common feedback (FB) words as the only vocal-verbal feedback part of a contribution, in relation to co-production with head movement, or not.

This suggests that feedback words not are longer when they are accompanied by head movements, than when they are not, at least not when considering head movements in general. However, when turning to more specific head movements, namely different nod types, a slightly different pattern emerge. Diagram 1 shows the mean duration of the five most common feedback words, as single feedback words, in relation to their accompaniment of single up nods, repeated up nods, repeated down nods, single down nods, and no nod at all.

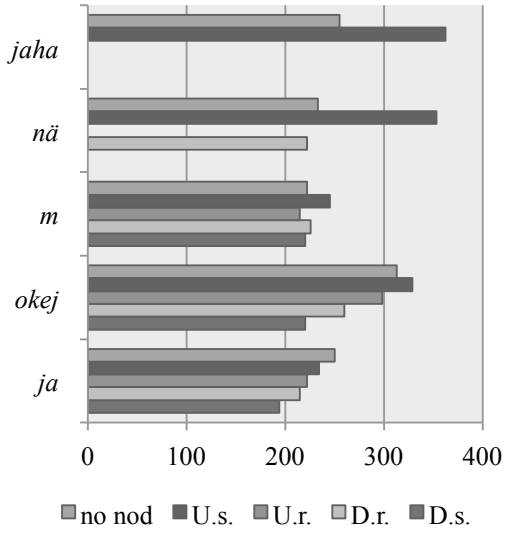

Diagram 1. Mean duration (ms) of the five most common single feedback words when co-produced with down-single (D.s.), down-repeated (D.r.), up-repeated (U.r.), up-single (U.s.), and no nod.

For all the words in Diagram 1, except for *ja* ('yes'), all have the longest mean duration when produced with single up nods. (It should be noted that *m* ('m') is uncommon with single up nods, see Table 4, so the mean duration of *m* ('m') with single up nod is based on only two instances.) So even though there is no consistent finding that words are longer when they are produced with head movements, than when they are not, it seems to be the case that feedback words are typically longer when they are produced with single up nods, than when they are not. That *jaha* ('yes, I see') have longer duration when produced with head movement (in general), see Table 5, should be understood in relation to its high co-occurrence with single up nods, see Table 4, and the observation that feedback words tend to be longer when produced with single up nods. Diagram 1 also shows a contrast between words accompanied by up-single nods and down single nods. All multimodal words are longest with up-single nods, while at least in the case of *ja* ('yes') and *okej* ('okay'), the shortest words are produced with down single nods. The duration of the word seems to vary systematically depending on what head movement accompanies the word. Another prosodic feature that also shows evidence of such systematic variation is the feature of pitch, which will be discussed below.

F0 contour	n	With head movement		Without head movement	
		n	%	n	%
Flat	183	116	63	67	37
Down	151	105	70	46	30
Up	123	51	41	72	59
Up-down	56	24	43	32	57
Complex-down	31	21	68	10	32
Down-up	26	14	54	12	46
Complex	32	21	66	11	34
Complex-up	16	13	81	3	19
Measuring error	55	24	44	31	56
Total	673	389		284	

Table 6. F0 contours of vocal-verbal feedback in relation to the accompaniment of head movement or not.

The most common F0 contours of the material are flat, down/falling and up/rising. These three types differ in their distribution with and without the accompaniment of head movement, see Table 6. Vocal-verbal feedback which have a flat or falling F0 contour are more common together

with head movement, than without, while vocal-verbal feedback having a rising F0 contour are more common without the accompaniment of head movement.

Looking at associations of different kinds of head movements and different prosodic features, we find a number of differences.

There seems to be a general trend that single upwards nods tend to co-occur with more stressed vocal-verbal feedback. Also repeated downward nods often co-occur with less stressed vocal-verbal feedback. This is shown in several prosodic dimensions.

Nod type	Freq. diff.	Freq. diff. ≥ 20%	Freq. dev.
Repeated down-nod	15%	21%	-3%
Single down nod	16%	37%	-6%
Repeated up nod	17%	28%	-2%
Single up nod	26%	54%	3%
None	18%	31%	12%

Table 7. Mean pitch measures of vocal-verbal feedback in relation to nod types. Freq. diff. ≥ 20% are the percentage of instances that have a frequency difference larger or equal to 20%.

Nod type	Mean duration of words (ms)
Repeated down-nod	223
Single down- nod	207
Repeated up-nod	239
Single up-nod	289
None	252

Table 8. Mean duration of single feedback words with different nods types (in milliseconds)

Nod type	Low saliency F0		High saliency F0	
	n	%	n	%
Repeated down-nod	104	76%	33	24%
Single down-nod	21	69%	12	31%
Repeated up-nod	67	69%	30	31%
Single up-nod	42	49%	44	51%
None	123	48%	131	52%

Table 9. Saliency of F0 contours of vocal-verbal feedback in relation to nod types. Low saliency F0 = flat, down, and complex-down. High saliency F0 = up, up-down, down-up, complex-up, and complex.

On average, vocal-verbal feedback co-occurring with single up nods have more prominent pitch features, such as more pitch variation (Frequency difference), in general higher pitch compared to mean speaker pitch (Frequency deviation). These single feedback words also have longer duration on average (Table 8), as well as a tendency to have more salient F0-curve characteristics (i.e. rising pitch at some point) (Table 9). As increased and/or rising pitch and increased duration are all considered to be typical prosodic features of stress, this suggests that the single up nod type is more likely to be co-produced with stressed vocal-verbal feedback than the other nod types are.

4 Summary and discussion

The results of this study are summarized below:

- Head nods are by far the most common type of head movement used for feedback (in Swedish first acquaintance conversations).
- In communicative feedback, words and nods are more frequently used in combination (multimodal), than on their own.
- Multimodal nods more frequently start before or at the same time as words, rather than words starting before nods.
- Nods have longer duration when produced with words (multimodal) than without, but words, however, are not typically longer when multimodal.
- Vocal-verbal feedback having a flat or falling F0 contour are more common together with nod, than without, while vocal-verbal feedback having a rising F0 contour are more common without the accompaniment of nod.
- Certain feedback words (*m* ('m') and *okej* ('okay')) are more often produced with nods, than others (*ja* ('yes') and *nä* ('no')).
- Furthermore, certain feedback words are typically associated with certain nod types, most prominently *okej* ('okay') and *jaha* ('yes, I see') with up nods, and *m* ('m') with repeated (down) nods.
- Single up-nods tend to occur with vocal-verbal feedback that have more salient prosodic features, while repeated down nods tend to occur with vocal-verbal feedback that have less salient prosodic features.

These results do to some extent differ from previous findings. First, Allwood and Cerrato (2003) found that feedback words were 20-40%

longer when produced with head movements, than without. This pattern was not confirmed for our data (see Diagram 1 and Table 8). (Note that nods are longer when they are multimodal, than when they are not.) Also, to be predicted from previous research on "visual prosody" is that emphasis in speech is likely to be produced with head movements (Graf et al, 2002; Swertz and Krahmer, 2010). Again, we do not find any unequivocal evidence for this pattern here. For example, feedback words do not typically have longer duration (Table 8), nor are they more salient (Table 9) when produced with nods, *per se*, than when they are produced without. Feedback words do have salient prosodic features and longer duration with some nod types, i.e. single up nods and to some extent repeated up nods, but not with other nod types, i.e. down nods. We therefore suggest that, how word and head movement are co-produced in multimodal feedback seems to be dependent on the *type* of word and the *type* of nod forming the multimodal contribution, rather than their co-production, *per se*.

Acknowledgments

This work is funded by the Swedish Research Council (VR) and the Nordic council (NOS-HS NORDCORP). We wish to thank Jens Allwood, Karl Johan Sandberg and Axel Olsson, as well as the anonymous reviewer for constructive criticism and suggestions.

References

Allwood, J. (1988) Om det svenska systemet för språklig återkoppling. In: P Linell, V. Adelswärd & P. A. Pettersson (ed.) *Svenskans beskrivning* 16, vol. 1. Linköping: Tema kommunikation, Linköpings universitet.

Allwood, J. and Cerrato, L. (2003) A Study of Gestural Feedback Expressions. *First Nordic Symposium on Multimodal Communication*. Paggio P. Jokinen, K. Jönsson, A. (eds). Copenhagen, 23-24 September 2003, pp. 7-22.

Allwood, J., Nivre, J. & Ahlsén, E. (1992) On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1): 1-26.

Allwood, J., Kopp, S., Grammer, K., Ahlsén, E., Oberzaucher, E. & Koppensteiner, M. (2007) The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41(3-4): 255-272.

Bull, P. and Connelly, G. (1985) Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3): 169-187.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S. Hralay, F. & Espesser, R. (1996) About the relationship between eyeborw movements and F0 variations. In: H. T. Bunnell & W. Idsardi *Procedings of ICSLP*, Philadelphia, PA, USA, pp. 2175-2178.

Graf, H. P., Cosatto, E., Strom, V. & Huang, F. J. (2002) Visual prosody: Facial movements accompanying speech. *Proceedings of the fifth IEEE International conference on automatic face and gesture recognition*.

Granström, B. and House, D. (2005) Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46: 473-484.

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

McNeill, D. (1985). So you think gesture are nonverbal? *Psychological Review, 92*, 350-371.

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2007) Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8-9):568-585.

Moubayed, S. A., Beskow, J. and Granström, B. (2010) Auditory visual prominence: from intelligibility to behaviour. *Journal on Multimodal User Interfaces*, 3(4): 299-311

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T. and Vatikiotis-Bateson, E. (2004) Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2): 133-137.

Scarborough, R., Keating, P., Mattys, S. L., taehong, C. and Alwan, A. (2009) Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, 51(2-3): 135-175.

Krahmer, E. and Swertz, M. (2009) Audiovisual prosody – introduction to the special issue. *Language and Speech*, 52(2-3): 129-133.

Swerts, M. and Krahmer, E. (2010) Visual prosody and newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, 38: 197-206.