Coordination of head movements and speech in first encounter dialogues

Patrizia Paggio

University of Copenhagen and University of Malta
paggio@hum.ku.dk, patrizia.paggio@um.edu.mt

Abstract

This paper presents an analysis of the temporal alignment between head movements and associated speech segments in the NOMCO corpus of first encounter dialogues [1]. Our results show that head movements tend to start slightly before the onset of the corresponding speech sequence and to end slightly after, but also that there are delays in both directions in the range of -/+ 1s. Various factors that may influence delay duration are investigated. Correlations are found between delay length and the duration of the speech sequences associated with the head movements. Effects due to the different head movement types are also discussed.
Index Terms: head movements, movement-speech alignment, delays, dialogues.

1. Background

Many studies have claimed that speech and gesture, in particular hand gestures, are two manifestations of the same underlying cognitive mechanism [2], [3], [4], [5], [6]. One aspect of this tight relation is the temporal coordination between the two modalities. It is generally agreed that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries [7], [8], [9], [10]. It has also been shown experimentally that subjects are sensible to asynchrony, especially when gesture strokes are made to lag behind the accompanying speech [11], and also that coordination with prosody contributes to the well-formedness of multimodal signals [12].

These studies deal with hand gestures, especially beats. Head movements often have the same quality of manual beats, by being rapid, simple and often repeated movements. Therefore, we would expect them also to show tight temporal synchronisation with the words they co-occur with. Coordination between head movements and speech is discussed in [13], where it is claimed that speakers' head movements are attuned to prosody in establishing peaks and prosodic boundaries especially in cases of high intensity. Furthermore, in [14], it is argued that coordination with speech, together with physical properties of head movements (cyclicity, amplitude, duration) are indicative of the diverse communicative functions of the movements themselves. However, the temporal synchronisation between the two modalities is not described in detail, and the datasets explored in these papers only consist of a couple of hundreds of head movements.

In this paper, we look at temporal synchronisation at the level of onsets and offsets of movements and associated speech, and we analyse a larger dataset.

2. The corpus

The data used in this study come from the Danish NOMCO corpus of first encounter dialogues, a collection of twelve video-recorded dialogues between Danish speakers for a total of about an hour of interaction. The annotation consists of the speech transcription as well as a rather fine-grained annotation of the speakers' gestural behaviour, including their head movements. In addition, each movement is explicitly linked to the speech segment which is semantically associated with it.

Figure 1: Annotation of a head movement in the Danish NOMCO corpus.

For instance, Figure 1 shows the ANVIL [15] annotation board concerning a head movement of type *jerk* (up-nod), which has been linked to the word *okay* in the speaker's own speech stream through the feature *MMRealationSelf*. More detail about the corpus, which is one of a collection of Nordic first encounter dialogues, can be found in [1], and [16].

3. Temporal coordination between head movements and speech

The total number of head movements in the NOMCO corpus is 3117. We are only interested in head movements that are linked to word sequences in the gesturer's own speech stream, and ignore unimodal head movements performed while the interlocutor is speaking. That leaves a subset of 2795 movements, which will be used to analyse movement-speech synchronisation in this study. The duration of most head movements in this dataset is around 1s, although there are occurrences of up to 7s (mean = 0.93s, sd = 0.58s). The duration of the word sequences linked with the head movements, on? the other hand, is on average shorter but with single outliers of 8s and 12s (mean = 0.59s, sd = 0.67s). The distributions are shown in Figure 2.

In what follows we analyse synchrony between head movements and associated speech sequences by looking at start and end delays between the two. A positive start delay means that

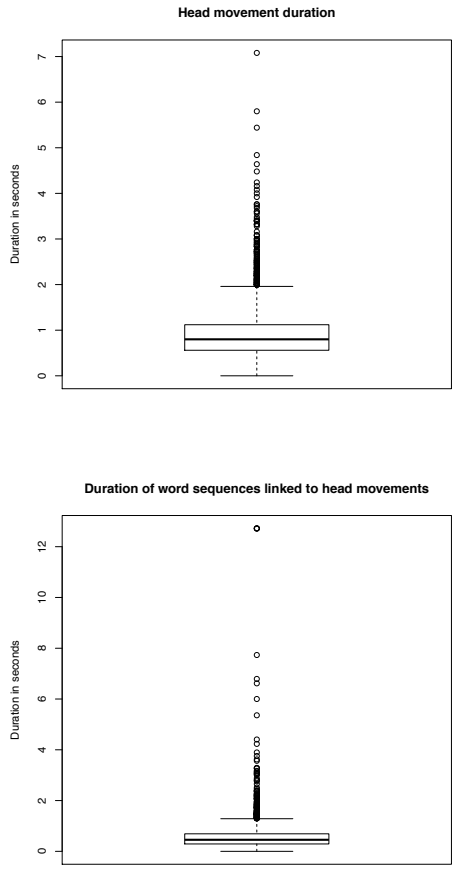

Figure 2: Distributions of the durations of head movements (above) and associated speech sequences (below).

speech onset follows movement onset, in other words that the head movement starts before the associated speech. A positive end delay, on the other hand, means that speech offset follows movement offset, in other words that the head movement ends before speech does.

On average, in our data head movements tend to start 0.05s before the onset of the associated speech sequence (sd = 0.40s), and to end 0.28s after its offset (sd = 0.64s). The histograms in Figure 3 show that in more than 2500 cases (out of the total 2795), delays range between -0.5 and 0.5, and that about 1750 delays are actually positive delays in the range 0 to 1, meaning that in almost two thirds of the cases head movements start before the corresponding speech. Looking at the end delays, on the other hand, we see that slightly more than 1800 are distributed in the range -1 to 0, meaning that in almost two thirds of the cases, the head movement ends up to 1s after speech offset. To have an intuition of what a one second delay means, we can compare it with the mean word duration in the whole NOMCO corpus, which is 0.21s, or the mean length of a linked speech sequence in the dataset, which is as we saw 0.59s. It can also be mentioned that in the already cited study in [11] it is found that subjects are sensible to asynchrony of as little as 0.2 seconds if a gesture lags behind speech, whereas in[12] it is claimed that subjects react to gesture-speech misalignments of at least 0.5 seconds. Thus, a delay of 1s is not negligible, in that it corresponds to four words, or two speech sequences.

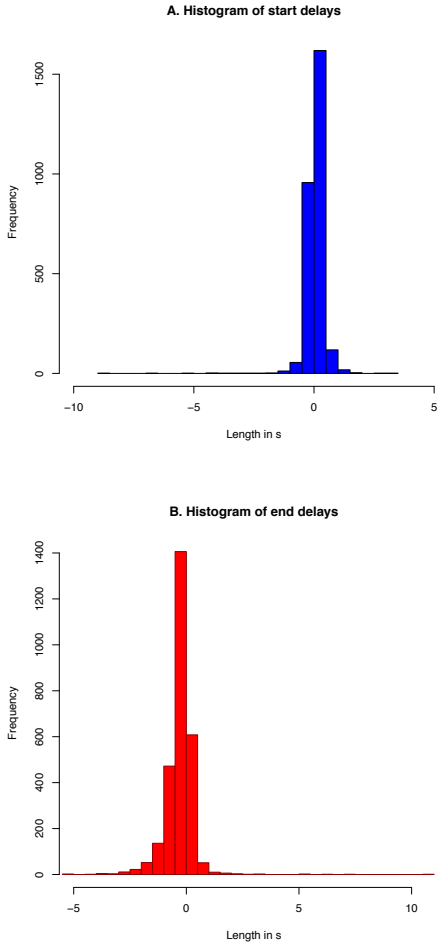

Figure 3: Start and end delays in the NOMCO corpus. In histogram A, bars to the left of zero (negative) correspond to speech preceding the onset of the corresponding movements, and those to the right to speech onset following movement onset. In histogram B, bars to the left of zero count speech ending before, and those to the right speech ending after movement offset. Histogram bins correspond to intervals of half a second.

In the remainder of the paper we will discuss a number of factors which may have an influence on the polarity and duration of the delays.

3.1. Delays in the individual conversations

Some variation can be observed in the individual conversations. In the top graph of Figure 4, which shows means and confidence intervals for start delay duration in the various files, we see that the mean delay duration varies from 0.13 (file M2_M4) to -0.06 (file M6_F1). All the means relating to end delay duration in

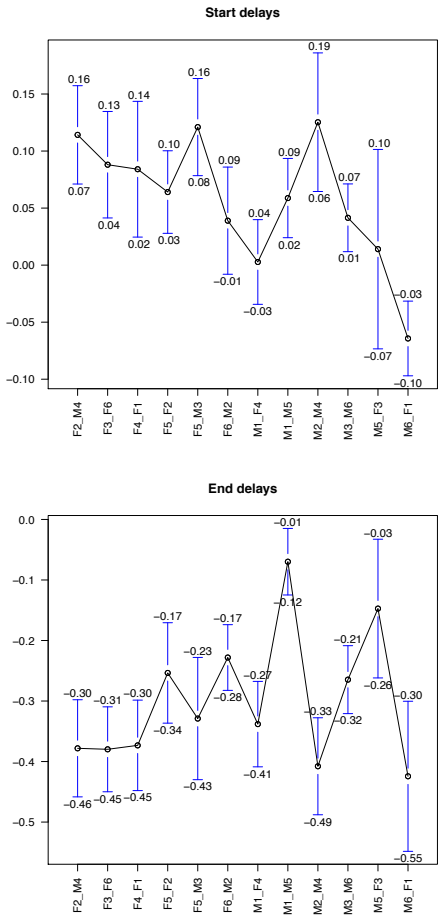

Figure 4: Duration of start and end delays in the twelve NOMCO conversations (means and confidence intervals).

the second plot of the same figure, on the other hand, are in the negative part of the chart (movement ending after speech) and vary from -0.07 (file M1_M5) to -0.42 (file M6_F1).

There is a very small but significant effect of conversation on start delay length (ANOVA, $F(11,2783) = 3.958$, $p<0.001$) and end delay length (ANOVA, $F(11,2783) = 7.387$, $p<0.001$). Only the differences between start delay duration in dialogue M6_F1 and seven of the other dialogues (F2_M4, F3_F6, F4_F1, F5_F2, F5_M3, M1_M5, and M2_M4) reach statistical significance (Tukey's HSD: p-values between < 0.05 and < 0.001). Looking at end delays, on the other hand, statistically significant differences are found between 14 of the pairwise comparisons, all of which involve either dialogue M1_M5 or M5_F3 (Tukey's HSD: p-values between < 0.05 and < 0.001).

3.2. Delays and individual speakers

We also looked at whether speakers differed from each other in their delay durations (Figure 5). We found a very small but significant effect of individual speaker on start delay length (ANOVA, $F(11,2783) = 2.633$, $p<0.001$) and end delay length (ANOVA, $F(11,2783) = 8.708$, $p<0.001$). As far as start delay duration is concerned, only the differences between speakers M4 and M1 on the one hand, and M4 and M6 on the other, reached significance, while 17 of the pairwise comparisons did when looking at end delays (Tukey's HSD).

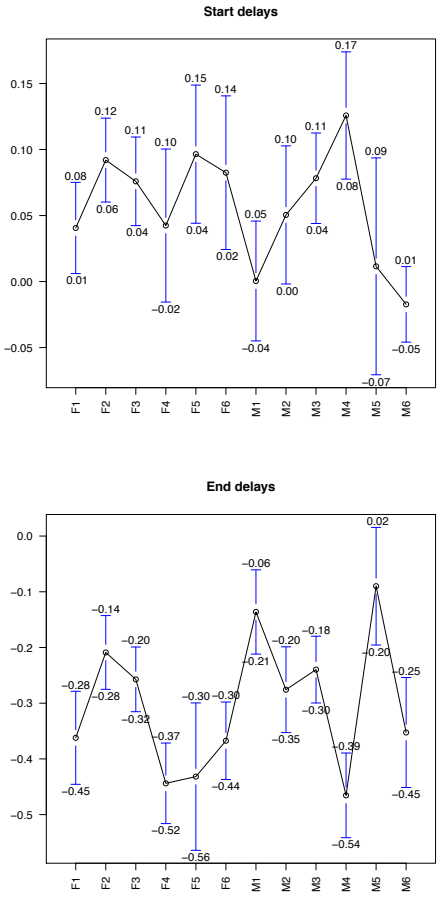

Figure 5: Duration of start and end delays produced the twelve NOMCO speakers (means and confidence intervals).

3.3. Delays and movement type

Start delay duration varies also depending on the type of movement (Figure 6), ranging on average between exact onset synchrony in the case of jerks (upnods) to an average positive delay of 0.12s in the case of waggles. None of the differences, however, reaches statistical significance. If we look at end delays, and leave out the category "Head Other", which collects cases of unclear movements, average duration varies between -0.13 for jerks to -0.54 for waggles. Head movement has in fact a small but significant effect on end delay duration (ANOVA, $F(8,2785) = 11.36$, $p<0.001$). The significant pairwise differences are shown in Table 1.

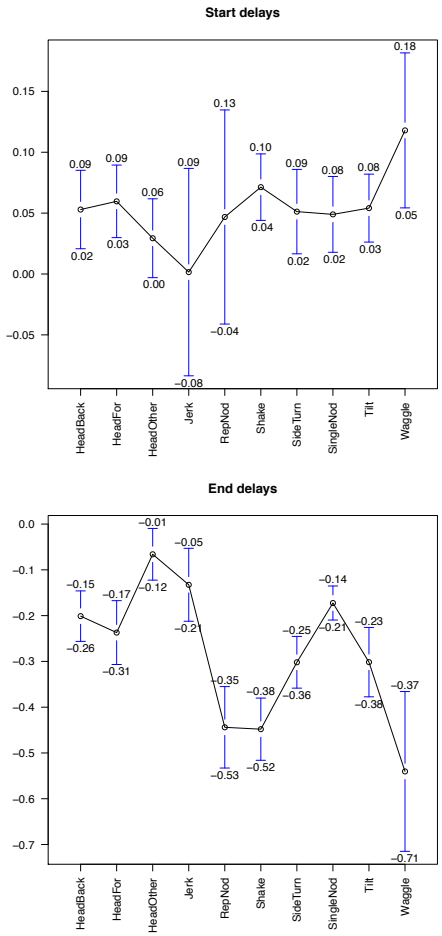

Figure 6: Start and end delays plotted against different head movement type (means and confidence intervals.)

3.4. Effect of movement and speech sequence duration on delays

Finally, the relations between delay length and head movement duration on the one hand, and between delay length and duration of associated speech segments on the other, were also investigated. In both cases, log values were used to diminish the effect of outliers on the correlation coefficient. No correlation was found between delay length and the duration of the head movements. On the other hand, a moderate *negative* correlation can be observed between start delay length and the duration of the speech segments (Pearson's r = -0.57), while a moderate *positive* correlation can be seen between end delay length and speech segment duration (Pearson's r = 0.40). The corresponding plots can be seen in Figure 7. In general, this means that the longer the speech chunk associated with the head movement is, the later the head movement starts and the earlier it ends. Interestingly, the strength of both correlations varies depending on head movement type, as shown in Tables 2 and 3.

We see that there are movement types for which there is a strong correlation between delay and linked speech duration both at onset and offset (jerks, repeated nods) and movement

Table 1: *Significant differences between end delay mean values for different head movement types (Tukey's HSD)*

Pairwise comparison	p value
RepeatedNod-HeadBackward	<0.001
RepeatedNod-HeadForward	<0.01
RepeatedNod-HeadOther	<0.001
RepeatedNod-Jerk	<0.001
Shake-HeadBackward	<0.001
Shake-HeadForward	<0.001
Shake-HeadOther	<0.001
Shake-Jerk	<0.001
SideTurn-HeadOther	<0.001
SingleNod-RepeatedNod	<0.001
SingleNod-Shake	<0.001
Tilt-HeadOther	<0.001
Waggle-HeadBackward	<0.01
Waggle-HeadForward	<0.01
Waggle-HeadOther	<0.001
Waggle-Jerk	<0.001
Waggle-SingleNod	<0.001
Waggle-Tilt	<0.05

Table 2: Pearson's r values showing correlation strength between start delay length and speech chunk duration related to different head movement types.

Head Movement Type	No. of cases	Pearson's r
Jerks	167	-0.94
Repeated nods	327	-0.73
Single nods	244	-0.52
Side turns	417	-0.50
Head other	199	-0.45
HeadB	237	-0.40
Tilts	455	-0.34
Shakes	325	-0.27
HeadF	338	-0.25
Waggles	86	-0.18
All	2795	-0.57

Table 3: Pearson's r values showing correlation strength between end delay length and speech chunk duration related to different head movement types.

Head Movement Type	No. of cases	Pearson's r
Jerks	167	0.81
Head other	199	0.65
Tilts	455	0.60
Repeated nods	327	0.57
Single nods	244	0.42
HeadF	338	0.36
HeadB	237	0.33
Side turns	417	0.25
Shakes	325	0.16
Waggles	86	0.03
All	2795	0.40

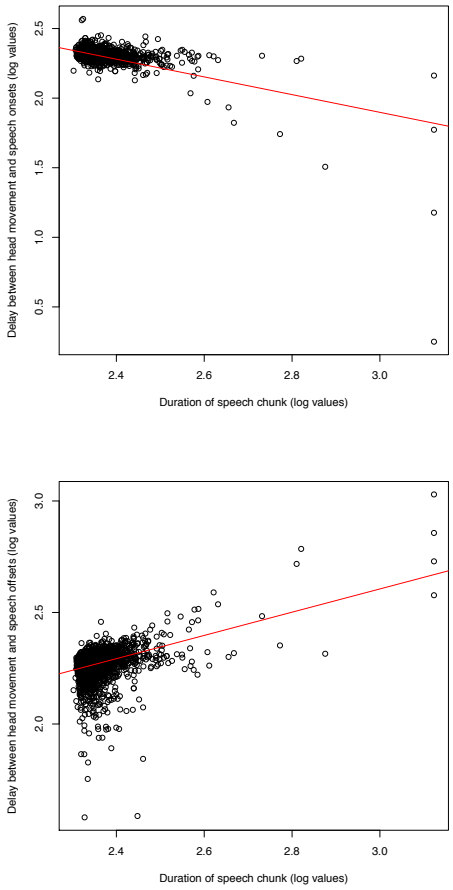

Figure 7: *Correlations between start delays and speech sequence duration (above), and between end delays and speech sequence duration (below). Log values are used.*

types for which the correlation is weak or non-existing at both ends (shakes, waggles). There are, however, also movement types that display a strong correlation between delay and linked speech duration only at onset (side turns, single nods), or only at offset (tilts)[1].

It is tempting to try to make sense of these differences in terms of the relative duration of the various movements, shown in Figure 8. Thus, jerks differ substantially from shakes and waggles both in terms of average duration and duration variance, for example, and they also behave differently in the correlations. Shakes resemble waggles, and likewise they behave very similarly as far as the correlations are concerned. On the other hand, side turns and tilts appear quite similar as far as duration is concerned, and yet they behave in different ways for what concern the correlation between delay and speech duration. In other words, although movement duration may have an effect on the way some head movement types and speech are

[1]The 'Head other' category also shows a strong correlation between end delay and speech offset. However, it is not clear what movement types are grouped in this class.

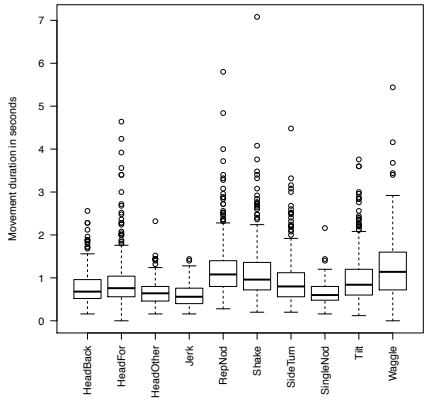

Figure 8: Head movement duration plotted against head movement type.

temporally coordinated, other factors are certainly at play. Examples might be the different kinetic properties of the various movement types, and the alignment between movement strokes and prosodic accents.

4. Conclusion

In a general sense it can be claimed that head movements are temporally synchronised with the associated speech sequences both at movement onset and offset. However, there are delays in both direction in the range of -/+ 1s, which is not a negligible time lag if we consider that subjects have been shown to be sensitive to delays of 0.5s.

Small effects on such variance may be explained in terms of conversation or speaker specific differences. Movement type also has a small effect on the offset delays, where we saw that especially shakes and waggles are responsible for significant differences with respect to other movement types. But the clearest feature that was found in this study was the correlation between delay length and duration of linked speech sequences, which is negative in the case of onset delays, and positive in the case of offset delays. In general, the longer the speech sequence is, the later the movement starts, and the earlier it finishes. This, in turn, can be interpreted as a general tendency for the overlap between head movement and speech sequence to be maximised.

This general pattern, however, varies depending on the movement type, with some types showing a more systematic adherence to the general tendency than others. While these differences seem to be related to the internal duration of the head movement in some cases (jerks, shakes, waggles), duration alone cannot explain the different behaviours of other movement types (e.g. nods, tilts and side turns). A more precise characterisation of the synchronisations patterns for these movement types probably needs to take into account the alignment between movement stroke and prosodic peak, or kinetic features such as amplitude and intensity.

We believe these results are interesting not only in their own right, but also in the context of development of speech production models involving different motional modalities.

5. References

[1] P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, and C. Navarretta, "The NOMCO multimodal nordic resource - goals and characteristics," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010.

[2] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago, 1992.

[3] ——, *Gesture and thought*. University of Chicago Press, 2005.

[4] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[5] S. Kita and A. Özyürek, "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking," *Journal of Memory and Language*, vol. 48, no. 1, pp. 16–32, 2003.

[6] J.-P. De Ruiter, "The production of gesture and speech," in *Language and Gesture*. Cambridge University Press, 2000.

[7] D. Bolinger, *Intonation and its parts: Melody in spoken English*. CA: Stanford: Stanford, 1986.

[8] A. Kendon, "Gesture and speech: two aspects of the process of utterance," in *Nonverbal Communication and Language*, M. R. Key, Ed. Mouton, 1980, pp. 207–227.

[9] D. P. Loehr, "Gesture and intonation," Ph.D. dissertation, Georgetown University, 2004.

[10] ——, "Aspects of rhythm in gesture and speech," *Gesture*, vol. 7, no. 2, 2007.

[11] T. Leonard and F. Cummins, "The temporal relation between beat gestures and speech," *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, 2010.

[12] G. Giorgolo and F. A. Verstraten, "Perception of 'speech-and-gesture' integration," in *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, 2008, pp. 31–36.

[13] U. Hadar, T. Steiner, E. C. Grant, and F. C. Rose, "Head movement correlates of juncture and stress at sentence level," *Language and Speech*, vol. 26, no. 2, pp. 117–129, 1983.

[14] U. Hadar, T. Steiner, and F. C. Rose, "The timing of shifts of head postures during conversation," *Human Movement Science*, vol. 3, no. 3, pp. 237–245, 1984.

[15] M. Kipp, *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com, 2004.

[16] C. Navarretta and P. Paggio, "Verbal and Non-Verbal Feedback in Different Types of Interactions," in *Proceedings of LREC 2012*, Istanbul Turkey, May 2012, pp. 2338–2342.