Temporal coordination of facial expressions and head movements in first encounter dialogues

Patrizia Paggio[*†], **Costanza Navarretta**[*]
[*]University of Copenhagen, [†]University of Malta
{paggio, costanza}@hum.ku.dk
patrizia.paggio@um.edu.mt

Abstract

This paper deals with the temporal coordination between facial expressions and co-occurring head movements in a multimodal corpus of first encounter conversations. In particular, we look at how the onset of facial expressions is coordinated with the first overlapping head movement, in other words which of the two modalities precedes the other and why. We find and discuss statistical main effects on the temporal delays between the two behaviours due to individual variation, type of head movement, and the communicative function of the multimodal signal. In particular, the analysis shows that when speakers give feedback, their facial expression becomes visible before the head starts to move, especially in the case of negative comments associated with frowning or scowling. The opposite is true when the multimodal signal is used as a comment to the speaker's own speech. The motivation for the analysis is to shed light on a less studied aspect of multimodal communication – an aspect that is relevant to the generation of natural multimodal expressions in ECAs.

Keywords: facial expressions, head movements, multimodal coordination

1. Background and goals

The coordination of different signals in human communication has been studied especially as regards gesture and speech, and there is considerable agreement that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries (Bolinger, 1986; Kendon, 1980; Loehr, 2004; Loehr, 2007). Experimental work has also clearly shown that people are sensitive to disruptions of the natural temporal alignment between the two modalities (Leonard and Cummins, 2010; Giorgolo and Verstraten, 2008). Coordination between head movements and speech, and how this is mediated by prosody, is discussed in Hadar et al. (1983) and (1984). More recently, Paggio (2016) and Paggio and Navarretta (2016) investigated the temporal alignment between head movements and co-occurring speech segments in multimodal data, and discussed a number of factors that affect the alignment.

Studies dealing with the relation between facial expressions and other expressive modalities have looked at the co-occurrence of several expression types. An early study found that children use eyebrow raises preceding head movements in connection with visual search (Jones and Konner, 1970). In a qualitative study, Kelner (1995) pointed out that enjoyment smiles co-occur with head movements towards the interlocutor while embarrassed smiles co-occur with head and gaze movements away from them (Keltner, 1995). Using quantitative methods, Cohnal *et al.* (2004b) studied correlations between lip-corner displacement in smiles and head or eye movements, and found that smile intensity correlates negatively with the presence of head movement in contexts involving embarrassment. Cohnal *et al.* (2004a) found that eyebrow raising is more likely to occur with forward head movements. Work where multimodal coordination of different expressions is used to model the behaviour of Embodied Conversational Agents (ECAs) include Cassell *et al.* (1999), Lee and Marsella (2006) and for emotional behaviours is discussed in Martin (2011). Finally, a study of how smiles and laughters can be generated based on the interlocutor's smiling and laughing behaviour, is in El Hadded *et al.* (2016).

In this paper, we focus on the coordination between facial expressions and head movements in cases in which there is indeed an overlap between the two modalities. In particular, we look at how the onset of facial expressions is coordinated with the first overlapping head movement, in other words which of the two modalities precedes the other and why. The motivation for the analysis is to shed light on a less studied aspect of multimodal communication – an aspect that is relevant to the generation of natural multimodal expressions in ECAs.

2. Multimodal facial expressions

The data for this study consist of 1448 facial expressions and 3117 head movements extracted from the Danish multimodal NOMCO corpus, an annotated collection of twelve first encounter dialogues involving six male and six female subjects of age 21 to 36. Each participant took part in a dialogue with a female and one with a male, for a total of about an hour of interaction. The two conversations took place on different days, and in both cases the dialogue participants had never seen each other before. The only instruction they received was to try to get to know each other. As a consequence, they spoke freely about a range of different topics. The dialogues were recorded in a studio, with the participants standing in front of each other, and were filmed by three cameras (Paggio and Navarretta, 2016).

The average duration of the facial expressions is 1.98s (sd=1.6). The spread of the duration is remarkable, with the shortest expression lasting 0.16s[1], and the longest 12.12. Smiles are the expressions showing the most variation in duration, with scowls showing the least. Head movements

[1]The expression is a short smile followed by a laughter. The annotators agreed about the two behaviours being separate expressions.

Table 1: Proportional conditional frequency of head movement types given co-occurring facial expression types

	Backward	Forward	HeadOther	Jerk	Nod	Shake	Turn	Tilt	Waggle	Sum
FaceOther	0.05	0.11	0.08	0.07	0.30	0.09	0.14	0.15	0.01	1
FrownScowl	0.10	0.13	0.11	0.02	0.14	0.12	0.18	0.16	0.04	1
Laughter	0.14	0.07	0.11	0.03	0.14	0.10	0.21	0.15	0.05	1
Raise	0.08	0.16	0.06	0.05	0.17	0.08	0.19	0.17	0.04	1
Smile	0.08	0.14	0.06	0.07	0.24	0.10	0.10	0.16	0.05	1

Table 2: Coordination of facial expression onsets with first co-occurring head movement: raw counts (proportions in parentheses)

Facial expression type	Before head	Same time as head	After head	Total
Smile	239 (.45)	35 (.06)	261 (.49)	535 (1)
Raise	122 (.40)	39 (.13)	146 (.47)	307 (1)
Laughter	63 (.45)	7 (.05)	69 (.50)	139 (1)
Frown/Scowl	39 (.38)	6 (.06)	58 (.56)	103 (1)
FaceOther	28 (.38)	6 (.08)	40 (.54)	74 (1)
Total	491 (.42)	93 (.08)	574 (.50)	1158 (1)

are shorter. Their mean duration is 0.93s (sd=0.58), with up-nods providing the shortest and least varying movements, and head shakes the longest outlier (7.08s). Head movements can be single or repeated. In our dataset there are 2315 single head movements, and 794 repeated ones. The mean duration for single movements is 0.82s (sd=0.48s), while it is 1.28 for repeated ones (sd=0.70s).

The majority of the facial expressions, i.e. 1158, or 80% of the total, co-occur with at least one head movement. Table 1 shows the proportion in which different types of head movements co-occur with the different kinds of facial expressions.

Of these, 491 (42%) start before, 93 (8%) at the same time, and 574 (50%) after the first co-occurring head movement. Frequency counts of the various facial expression types against their onset relation with the first co-occurring head movement are shown in table 2. In general, it can be concluded that there is a very high likelihood for facial expressions to be accompanied by head movements. However, whether the onset of the facial expression precedes or follows the onset of the head movement is equally likely. Nevertheless, a χ-squared test of independence showed that the type of onset delay depends on the facial expression type (χ^2=15.87, df=8, p-value=0.04429s). This dependency is mostly due to the fact that eyebrow raises (*Raise*) tend to start at the same time as the co-occurring head movement proportionally more often than the other types, while frowns and scowls (*Frown / Scowl*) tend to start after the onset of the head movement more often than the other types. There is also a slight tendency for *Smile* and *Laughter* to start before the head movement more often than expected.

These differences may well be due to different physical characteristics of the signals. For instance, eyebrow movements are quite small and their onset may therefore be more tightly coordinated with that of short accompanying head movements such as nods and turns. Conversely, smiles and laughters may imply a longer preparation phase and therefore tend to start earlier than the accompanying head movement.

3. Temporal coordination

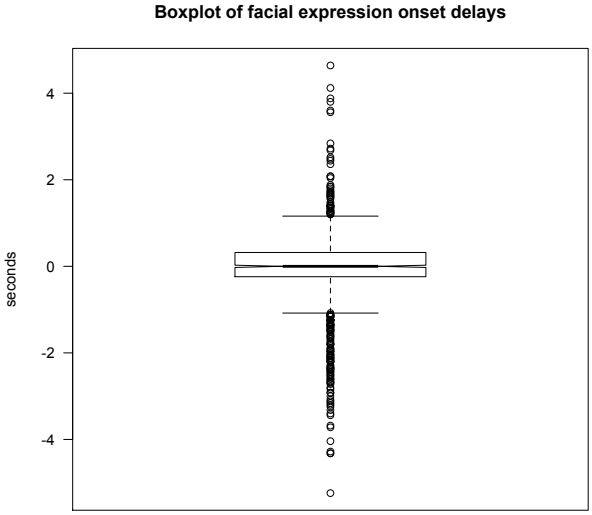

Figure 1: Boxplot of the distribution of onset delays between facial expressions and the first overlapping head movement. Positive delays indicate facial expressions starting after the co-occurring head movement.

In this section we look at the temporal coordination between the two co-occurring behaviours in a more fine-grained way. The mean onset delay between the two modalities is -0.05s (sd=0.9), indicating that the behaviours on average are almost coincidental (with a tiny likelihood for the face starting to move before the head), but that there is also considerable variation. The plot in figure 1 shows the distribution of the duration of the onset delays between facial expressions and the first overlapping head movement. Most of the delays are in the area between –1s (facial expression starting before the onset of the head movement), and +1s (facial expression starting after the onset of the head movement). There are, however, quite a number of outliers in both negative and positive ranges so that the data do not conform to the normal distribution (Shapiro-Wilk normality test, W=0.85783, $p < 0.001$).

Statistical tests show a main effect of individual speaker variation (Kruskal-Wallis: χ^2=44.002, df=11, p-value<0.001), an effect of head movement type (Kruskal-

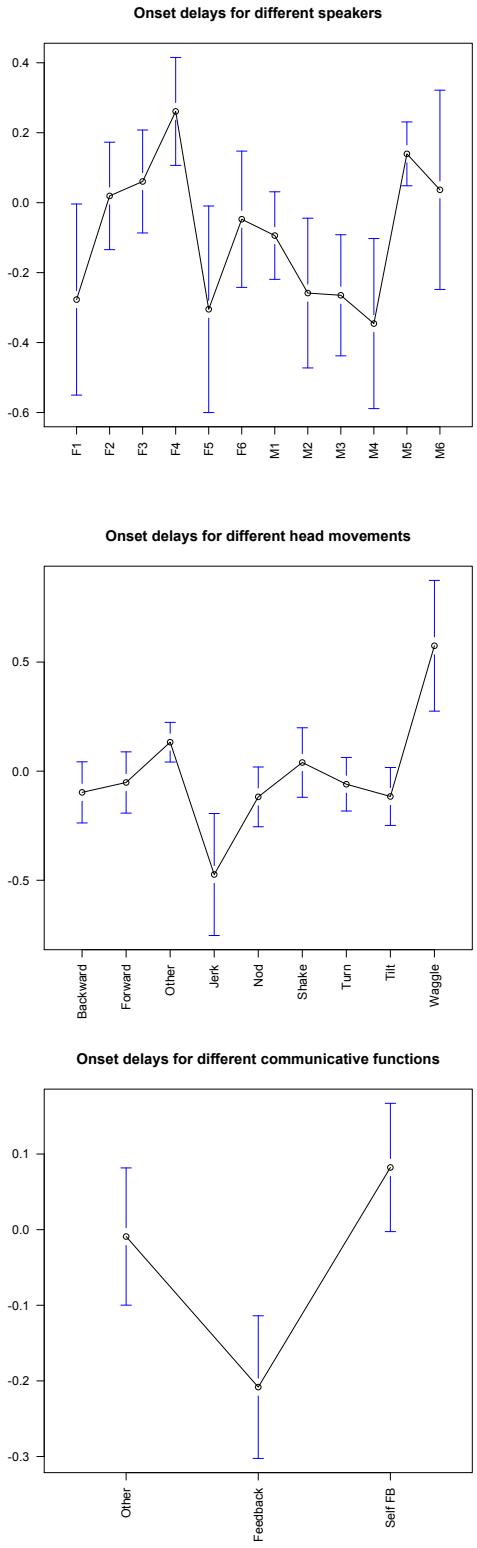

Figure 2: Mean values and confidence intervals for the temporal coordination between onsets of facial expressions and co-occurring head movements according to individual speakers (plot on top), associated head movement (plot in the middle), and function of the signal (plot below). Positive values indicate that facial expressions start after the onset of the head movement.

Wallis: χ^2=39.689, df=8, p-value<0.001), and an effect of function (Kruskal-Wallis: χ^2=22.802, df=3, p-value<0.001) on the distribution of the start delays. The effect of facial expression type, on the contrary, does not reach significance in spite of the results of the χ^2 test on the figures in table 2.

As can be seen from the topmost plot showing mean values and confidence intervals for different speakers in figure 2, only four of the speakers (F2, F3, F6 and M1) display an average delay around 0s, whilst the rest of the speakers have either a positive or a negative mean delay onset. Most of the significant differences involve F4 and M5[2]. As for the head movement type (middle plot in figure 2), negative delays are seen especially together with *Jerk* (up-nod) and positive ones with *Waggle*. Up-nods imply a backward movement of the neck which may physically be slightly more demanding than a forward movement, and have the effect of the movement becoming visible after the onset of the co-occurring facial expression. Conversely, waggles tend to precede the associated facial expressions. Waggles are rather complex and relatively long on average (mean duration=1.2s), characteristics which may explain why they are initiated early in the multimodal contribution. Most of the significantly different pairwise comparisons predictably involve *Waggle*, but the comparisons between delays involving *Jerk* and *Other* as well as *Jerk* and *Shake* also show a significant difference. This is not surprising since shakes are similar to waggles in being complex movements in which the head moves repeatedly in different directions.

A particularly interesting effect on the temporal coordination between facial expressions and head movements is the one relating to the communicative function assigned to the multimodal signal. Such dependence could in fact be exploited in the generation of facial expressions and head movements in ECAs. In this paper we distinguish between three function types: *CPU*, which stands for Contact, Perception and Understanding, for signals eliciting or giving feedback; *Self Feedback* for signals used by the speaker to comment their own contributions; and any other function[3]. The lowest plot in figure 2 shows that feedback to others and self feedback behave quite differently, with self feedback signals displaying a delay of about 1s on average, and feedback signals showing delays in the other end of the scale (about -0.2s on average) . In other words, when speakers react to their own speech, they tend to move the head first. When they give feedback, they tend to move the face first. This difference is statistically significant.

Finally, in the plot in figure 3 we show the combined effect

[2]All pairwise comparisons after the Kruskal-Wallis tests were done using the Dunn test with the Benjamini-Hochberg p-value adjustment method.

[3]Other functions relate to turn taking, discourse structuring, information stucture, etc. as defined in the MUMIN coding scheme (Allwood et al., 2007). Note that some of the functional categories in our annotations have a direct correspondence with discourse act categories in the ISO 24617-2 standard (`https://www.iso.org/standard/51967.html`). This is for example the case for the Auto- and Allo-Feedback dialogue acts, which have the same semantics as the MUMIN's SelfFeedback and FeedbackGiving attributes.

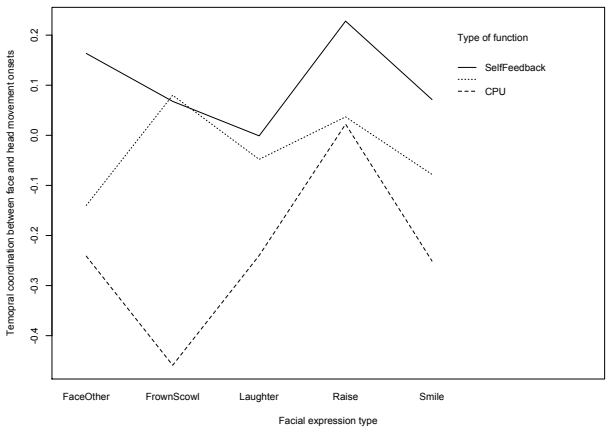

Figure 3: Interaction plot showing the combined effect of communicative function and facial expression type on the temporal coordination between facial expressions and co-occurring head movements.

of function and facial expression type. We see that the tendency for feedback behaviours (CPU in the figure) appearing in the negative end is stronger in the case of frowns and scowls, whereas in the case of eyebrow raises the different functions do not affect the direction of the delay much.

4. Discussion and conclusion

In general, our data clearly show that facial expressions have a strong tendency to co-occur with head movements and to be aligned with them at the onset. There are, however, delays in both directions. We have found interesting patterns concerning how the delays are distributed depending on the facial expression type. Thus, the onset of eyebrow raises is more tightly coordinated with the onset of the first co-occurring head movement, whereas both smiles and laughters tend to be initiated slightly earlier. These differences, however, do not reach significance in our data.

Significant effects on the temporal coordination between co-occurring behaviours in the two modalities, on the contrary, were found for head movement type and function of the signal in addition to individual variation. The effect due to head movement type can be explained at least partially in terms of the physical characteristics of the movements, with complex movements such as waggles showing a tendency to be initiated before the co-occurring facial expression. More interestingly, whether the onset of a facial expression (slightly) precedes or follows the onset of the first co-occurring head movement also depends on the function of the multimodal behaviour. In particular, we have found that when speakers give feedback, their facial expression becomes visible before the head starts to move, especially in the case of negative comments associated with frowning or scowling. Conversely, when the multimodal signal is used as a comment to the speakers' own speech contribution, the head movement tends to be noticed first. Since facial expressions are one of the strongest signals of attitudinal and emotional states, these results seem to indicate that in the case of a comment to the interlocutor's contributions, facial reactions are more immediate than feedback expressed by movements of the head. Head movements and facial expressions in our data have the same communicative function, that is were reinforcing each other, or have a function of repetition using the terminology by Poggi and Caldognetto (1996). The temporal relation between the two behaviours in other cases, for instance contradiction, should be investigated in different data.

Interactions between the various variables involved in our analysis are difficult to test statistically because of the non-normal distribution of the data, and were only illustrated graphically in this paper. In future, we intend to explore such interactions by applying machine learning techniques to the problem of predicting the alignment between facial expressions and head movements from the formal and functional factors discussed in this study. Linear mixed effects models could also be applied to investigate the interactions between the various factors.

To test the generality of our findings, it would be interesting to conduct similar analyses using data from different communicative situations as well as produced by speakers from different cultural backgrounds. We would also be interested in verifying if other patterns of behaviour than those found in the corpus would seem unnatural when implemented in an ECA.

A relevant and interesting issue we have not investigated, is how facial expressions are structured internally, and whether they contain a phase comparable to the stroke in hand gestures, see e.g. Kipp (2004). If or when they do, it is reasonable to assume that the onsets of facial expressions and head movements will be coordinated in such a way as to ensure that the strokes of the two behaviours are aligned. A related issue also not dealt with here is what happens when a protracted facial expression – for example a smile – overlaps with several distinct head movements. The way in which the temporal coordination between the two modalities should be described in such cases is far from clear, and will be left for future research.

5. Acknowledgements

We would like to thank the three reviewers for their useful comments.

6. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Jean-Claude Martin, et al., editors, *Multimodal Corpora for Modelling Human Multimodal Behaviour*, volume 41 of *Special issue of the International Journal of Language Resources and Evaluation*, pages 273–287. Springer.

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

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 520–527.

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J., and Moriyama, T. (2004a). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 1, pages 610–616, Oct.

Cohn, J. F., Reed, L. I., Moriyama, T., Xiao, J., Schmidt, K., and Ambadar, Z. (2004b). Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 129–135, May.

El Haddad, K., Çakmak, H., Gilmartin, E., Dupont, S., and Dutoit, T. (2016). Towards a listening agent: A system generating audiovisual laughs and smiles to show interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 248–255, New York, NY, USA. ACM.

Giorgolo, G. and Verstraten, F. A. (2008). Perception of 'speech-and-gesture' integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, pages 31–36.

Hadar, U., Steiner, T., Grant, E. C., and Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(2):117–129.

Hadar, U., Steiner, T., and Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, 3(3):237–245.

Jones, N. G. B. and Konner, M. (1970). An experiment on eyebrow-raising and visual searching in children. *Journal of Child Psychology and Psychiatry*, 11(4):233–240.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality & Social Psychology*, 68:441–454.

Kendon, A. (1980). Gesture and speech: two aspects of the process of utterance. In M: R. Key, editor, *Nonverbal Communication and Language*, pages 207–227. Mouton.

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

Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer.

Leonard, T. and Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10):1457–1471.

Loehr, D. P. (2004). *Gesture and Intonation*. Ph.D. thesis, Georgetown University.

Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2).

Martin, J.-C., Devillers, L., Raouzaiou, A., Caridakis, G., Ruttkay, Z., Pelachaud, C., Mancini, M., Niewiadomski, R., Pirker, H., Krenn, B., Poggi, I., Caldognetto, E. M., Cavicchio, F., Merola, G., Rojas, A. G., Vexo, F., Thalmann, D., Egges, A., and Magnenat-Thalmann, N., (2011). *Coordinating the Generation of Signs in Multiple Modalities in an Affective Agent*, pages 349–367. Springer Berlin Heidelberg, Berlin, Heidelberg.

Paggio, P. and Navarretta, C. (2016). The Danish NOMCO corpus multimodal interaction in first acquaintance conversations. *Journal of Language Resources and Evaluation*, pages 1–32.

Paggio, P. (2016). Coordination of head movements and speech in first encounter dialogues. In *Proceedings of the 3rd European Symposium on Multimodal Communication*, Linköping Electronic Conference Proceedings, pages 69–74.

Poggi, I. and Caldognetto, E. M. (1996). A score for the analysis of gestures in multimodal communication. In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech Applied Science and Engineering Laboratories*, pages 235–244.