Aspects of co-occurring syllables and head nods in spontaneous dialogue

Simon Alexanderson, David House, Jonas Beskow

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
simonal@kth.se, davidh@speech.kth.se, beskow@speech.kth.se

Abstract

This paper reports on the extraction and analysis of head nods taken from motion capture data of spontaneous dialogue in Swedish. The head nods were extracted automatically and then manually classified in terms of gestures having a beat function or multifunctional gestures. Prosodic features were extracted from syllables co-occurring with the beat gestures. While the peak rotation of the nod is on average aligned with the stressed syllable, the results show considerable variation in fine temporal synchronization. The syllables co-occurring with the gestures generally show greater intensity, higher F0, and greater F0 range when compared to the mean across the entire dialogue. A functional analysis shows that the majority of the syllables belong to words bearing a focal accent.

Index Terms: Gestures, prosody, motion capture, beats, head nods, stressed syllable

1. Introduction

Along with intonation there are a wealth of visual gestures, including head, facial and body movements, which co-occur with speech adding emphasis and prominence to portions of the utterance and contributing to the flow of the dialogue. Beat gestures such as rapid hand movement as described by [1] are particularly interesting in this regard as they coincide and appear to be synchronized with prosodic and intonational peaks related to prominence. They also share the same prominence-lending function as the focal or sentence accent, but they can be repetitive, marking the stress or rhythmical structure of an utterance.

In terms of timing and synchronization, there are many similarities between intonational gestures and visual gestures produced in accompaniment with speech. Both intonation and visual gestures are free to vary across the vowels and consonants of the segments. In intonation, however, this variation is restricted by the specific patterns used by a language to signal meaning in spoken interaction.

If we wish to study the timing of gestures in the same way as we approach timing in intonation, we currently lack an established methodology to extract and analyze gestures, especially gestures occurring in spontaneous dialogue. The Spontal corpus of Swedish dialogue provides a rich database as a point of departure for testing gesture extraction and analysis methodology. The database, containing more than 60 hours of unrestricted conversation in over 120 dialogues between pairs of speakers is comprised of high-quality audio and video recordings (high definition) and motion capture for body and head movements for all recordings [2]. Motion capture offers many possibilities in gesture research. Compared to video-based head pose estimators, it offers increased accuracy and higher frame-rate (100 fps in our corpus).

Moreover, manual annotation of head-gestures in spontaneous dialogue involves a number of difficulties. Most speakers move their heads extensively while speaking and many gestures may be subtle, or involve simultaneous rotations around several axes. Among the sources of disagreement are boundaries and location of maximum extent of gesture segments. In order to facilitate and accelerate the processing and annotation of data such as is represented in the Spontal corpus, we are developing new semi-automatic methods overcoming some of these problems.

In this paper we report on aspects of synchronization and timing between the head gestures annotated as having a beat function and the co-occurring syllables having focal accent or stress. We also present some data concerning prosodic aspects of the co-occurring syllables.

2. Method

We employed a semi-automatic annotation procedure consisting of two steps. First an automatic head-gesture segmentation algorithm is applied to the motion capture data, generating candidate head gesture segments. Then the segments are manually classified by the annotators. In this section we give a brief overview of the procedure, see [3] for an in-depth description.

In addition to gesture annotation we also processed the audio files of the speakers and generated pitch and intensity data, talk vs. no-talk segmentation and syllable segmentation. This was done to be able to investigate the relationship between the head-gestures and prosodic features.

In this exploratory stage we chose one of the dialogues from the Spontal database where the speakers demonstrated a relatively large number and variety of head-movements. The participants were a male and female who did not know each other.

2.1. Automatic segmentation of head-nods

A simplistic segmentation approach was used for head-nod segmentation. The Euler angles of the head motion were calculated from the motion capture data using standard rigid body fitting algorithms. During a head nod the pitch angular velocity follows an oscillatory movement during a short time period. The automatic segmentation algorithm places segment boundaries at a local maximum followed by a minimum (or vice versa) of the pitch angular velocity, provided that their separation in time is below a specified threshold (1000 ms was empirically set for our data). A more detailed description can be found in [3].

2.2. Manual classification

After running the automatic processing, the resulting candidate segments were examined and manually classified by two

annotators. To carry out the classification, the annotators viewed each segment in a specially designed 3D player which plays time-synchronized audio and displays 3D point data of the motion capture markers together with animated characters following the 3D marker motion, see figure 1. This was done to enable the annotators to work on the same signal as the algorithm. In addition, the 3D-player allows for alternate view-points of the scene and maintains the high frame rate of the motion data. In some uncertain cases the annotators also used the video data as a reference. In future releases we plan to include this feature in the player.

Figure1: *3D viewer for manual annotation.*

As expected, the segments from the automatic process did not only contain unambiguous beat gestures, but also gestures with other functions co-occurring with speech [4]. Such other functions were feedback, confirmation, word or phrase intensification and listing of lexical items. Moreover, some of the extracted gestures did not appear to co-occur with a stressed syllable.

Therefore, an annotation scheme was devised with two main queries: Q1, "Is there a clear nod in synchrony with a stressed syllable?" and Q2, "Is the nod multifunctional?" If the answer to the first query is positive the second query is also answered. This scheme resulted in three categories: 1. No clear nod in synchrony (no sync), 2. A clear nod with a beat function (beat-function), and 3. A clear nod which is multifunctional (multi-function). The annotation time was approximately two nods per minute.

2.3. Prosodic features

The pitch and intensity curves were extracted from the audio signals of the speakers using the SNACK toolkit [5]. Also syllable boundaries and nuclei were derived by applying Mermelstien's convex-hull algorithm [6].

After gesture- and prosodic feature extraction was performed, we annotated the function of the stressed syllables co-occurring with the subset of the female speaker's nods which were marked as having a beat function by both annotators. Only the gestures from the female speaker were analyzed due to the small number of beat gestures annotated for the male speaker (see section 3.1).

The nods were displayed and time-aligned with a speech spectrogram display using WaveSurfer [5]. Based on an auditory and visual analysis, one annotator, a trained phonetician, marked the primary stressed syllable of the word co-occurring with each head nod. The syllable was also annotated as to whether or not it comprised a word or was part of a word containing a focal accent. If there was no stressed syllable occurring simultaneously with the head nod, the nearest stressed syllable was marked and annotated.

3. Results

3.1. Gesture extraction

The automatic segmentation algorithm was applied on the 20 minute dialog, extracting 64 nod-segments for speaker 1 (male) and 150 segments from speaker 2 (female). The manual classification by the two annotators resulted in a 69% and 65% agreement for the head-nods of speaker 1 and speaker 2 respectively. As is displayed in table 1 the annotators showed greatest agreement for the category with no syllable-synchronization. Less agreement was obtained from the categories beat-function and multi-function. Annotator 1 perceived more of the nods as having a pure beat-function while annotator 2 perceived more as being multi-functional.

Table 1: *Results of the manual annotation showing class and agreement. The first number in each pair is the male speaker, the second is the female.*

Class	Annotator 1	Annotator 2	Agreement
No sync	23/44	15/41	13/29
Beat-function	9/74	11/66	4/50
Multi-function	32/32	38/43	27/18
Total	64/150	64/150	44/97

3.2. Syllable annotation

The results of the annotation of the stressed syllables co-occurring with the 50 head nods are shown in table 2. 36 of the 50 syllables comprised or were part of a word bearing a focal accent. 14 syllables which contained primary stress did not bear focal accent nor were they part of a word bearing focal accent.

Table 2: *Results of syllable annotation.*

Class	Number
Stress-focus	36
Stress-no focus	14
Total	50

All but three of the 50 syllables overlapped in time with the head nod. In one of these cases of non-overlap, the head nod overlapped with the following word. In the second case, the nod overlapped with the syllable bearing secondary stress of the focally accented word. In two cases the co-occurring stressed syllable was an interjection and was annotated as not containing a focal accent. In nine cases, the convex-hull algorithm included more than one single syllable. No correction for this was made in the context of this study.

3.3. Gesture timing related to syllables

The subset of the 50 gestures annotated as having a beat function for the female speaker was also analyzed in terms of timing related to its co-occurring stressed syllable. Figure 3 shows the time difference between two different anchor-points of the syllable (onset and nucleus) and three different phases of the nod: peak angular velocity of the downward phase (p1), max rotation (p2) and peak angular velocity of the upward phase (p3). The timing relationship between the gesture and the syllable does not seem to be influenced by the choice of syllable anchor-point. The timing relationships show a considerable amount of variation regarding the question of gesture synchronization with the syllable. However, as can be seen in figure 3, the peak rotation of the nod (p2) is on average aligned with the nucleus of the stressed syllable.

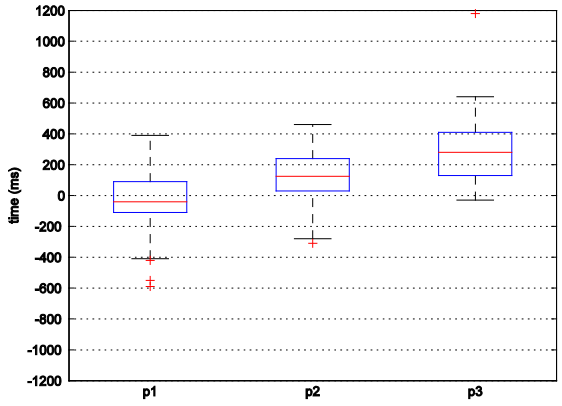

Figure 2: *Timing of the ONSET of the co-occurring stressed syllable with respect to three different phases of the nod: peak angular velocity of the downward phase (p1), max rotation (p2) and peak angular velocity of the upward phase (p3).*

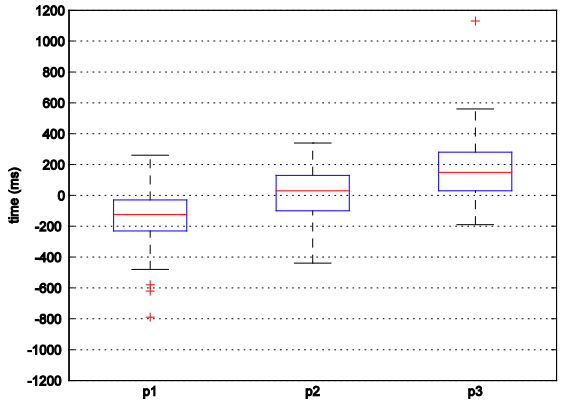

Figure 3: *Timing of the NUCLEUS of the co-occurring stressed syllable with respect to three different phases of the nod: peak angular velocity of the downward phase (p1), max rotation (p2) and peak angular velocity of the upward phase (p3).*

3.4. Prosodic features

When compared with mean values across the total dialogue, the stressed syllables showed greater integrated intensity, higher F0 at the nucleus, and greater F0 range.

Table 2: *Comparison of syllable features.*

	integrated intensity (dBs)	F0 at nucleus (Hz)	F0 range (Hz)
Mean stressed syllable	14.9	214	87
Mean across the total dialogue	11.1	180	67

4. Discussion

In this paper we have made use of a new semi-automatic process of extracting and annotating head nods in spontaneous dialogue. We have examined three aspects of the characteristics of the annotated beat nods and their co-occurring stressed syllables: temporal alignment, prosodic features, and the prosodic function of the syllable in the utterance. The results must be seen as quite preliminary due to the relative small sample.

The results of the analysis of the timing between beat nods and syllables show a general alignment in time between the maximum extent of the nod and its corresponding syllable, but with a considerable variation. This is consistent with the results reported in [7] for hand and arm beat gestures. Some more extreme values are seen for our head nod data which can reflect a greater temporal variability for head nods in spontaneous speech than for hand and arm beat gestures in read speech. Similar results regarding alignment between eyebrow raises and pitch accented syllables in English have been presented in [8], but here again the results show considerable variation in the timing of the gestures.

The analysis of the prosodic features clearly shows that the syllables co-occurring with the beat nods convey acoustic prosodic prominence. Here we see how gestures and prosody work together to add prominence in the spontaneous dialogue. Furthermore, the functional analysis of the co-occurring syllables shows that a clear majority of them belong to words bearing focal accent (phrasal prominence). In disyllabic or compound words with Swedish word accent 2, the nods generally coincide with the primary stressed syllable, although there are cases where the nod coincides with the physical focal accent on the syllable with secondary stress. Here again we see the beat nod contributing to the function of prominence in the dialogue. More speakers in other dialogues in the Spontal corpus are being analyzed to determine how general these findings are.

Moreover, the methods used in this study can also be seen as a starting point for further work in the field of automatic recognition and classification of multimodal communication. While head gestures, and in particular relatively small gestures, have been problematic for manual annotation schemes, the semi-automatic method tested here shows promise for the selection and fast annotation of multimodal data. Given the fact that non-verbal communication and verbal communication are tightly coupled, the motion data may provide important and robust

features for machine-learning techniques. In this study we started an investigation along this path by analyzing features that may be useful for prominence detection. This method may also prove useful in analyzing features related to other communicative functions such as feedback and turn-taking. More available data and the development of automatic methods and tools should better enable us to compare and evaluate results such as these.

5. Acknowledgements

The work reported here is carried out within the projects: "Timing of intonation and gestures in spoken communication," (P12-0634:1) funded by the Bank of Sweden Tercentenary Foundation, and "Large-scale massively multimodal modelling of non-verbal behaviour in spontaneous dialogue," (VR 2010-4646) funded by the Swedish Research Council.

6. References

[1] McNeill, D., Gesture and thought. Chicago: The University of Chicago Press, 2005.

[2] Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S. and House, D., "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture", in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner and D. Tapias [Eds], Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valetta, Malta: 2992−2995, 2010.

[3] Alexanderson, S., House, D. and Beskow, J., "Extracting and analyzing head movements accompanying spontaneous dialogue", in Proc. of Tilburg Gesture Research Meeting, Tilburg, 2013.

[4] McClave, E., "Linguistic functions of head movements in the context of speech", Journal of Pragmatics 32: 855−878, 2000.

[5] Sjölander, K. and Beskow, J., "WaveSurfer - an open source speech tool", in B. Yuan, T. Huang & X. Tang [Eds], Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing. Beijing: 464−467, 2000.

[6] Mermelstien, P., "Automatic segmentation of speech into syllabic units", Journal of the Acoustical Society of America 58: 880−883, 1975.

[7] Leonard, T. and Cummins, F., "The temporal relation between beat gestures and speech", Language and Cognitive Processes 26: 1457−1471, 2011.

[8] Flecha-Garcia, M. L., "Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English", in Proceedings of CogSci-2007. Austin, Texas: 1753, 2007.