Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016)

Copenhagen, 29-30 September 2016

Editors:

Patrizia Paggio[1, 2] and Costanza Navarretta[1]

[1]University of Copenhagen, [2]University of Malta

Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication
(MMSYM 2016), Copenhagen, 29-30 September 2016
Edited by Patrizia Paggio and Costanza Navarretta
ISBN: 978-91-7685-423-5

Linköping Electronic Conference Proceedings No. 141
ISSN: 1650-3686, eISSN: 1650-3740
URL: http://www.ep.liu.se/ecp/contents.asp?issue=141

Content

Introduction

This volume presents a selection of the papers presented at MMSYM 2016, the 4th European and 7th Nordic Symposium on Multimodal Communication. The symposium aimed to provide a multidisciplinary forum for researchers from different disciplines who study multimodality in human communication as well as in human-computer interaction. It was organised by the Centre for Language Technology (CST), University of Copenhagen, and took place at the University of Copenhagen on September 29-30, 2016. The invited speaker was Adam Kendon which discussed the relation of language and gesture in communication.

The symposium followed up on a tradition established by the Swedish Symposia on Multimodal Communication held from 1997 till 2000, and continued by the Nordic Symposia on Multimodal Communication held from 2003 to 2012. Since 2013 the symposium has acquired a broader European dimension, with editions held in Malta, Estonia and Ireland. In 2016 the symposium returned to the University of Copenhagen.

As in the preceding editions, we aimed to present a broad view of the field by accepting papers on a wide range of topics and employing different methodologies in order to reflect the largely inter-disciplinary nature of this research filed. This breadth is also maintained in the choice of papers presented here, which span from modelling and automatic detection of different kinds of communicative movement to the use of non-verbal signals for specific purposes in different communicative situations and different languages. The methods adopted also vary a great deal, encompassing quantitative analyses, experimental and neurocognitive investigations, and semiotic taxonomies. The individual contributions are briefly summarised below, and positioned with respect to this range of topics and methods.

Frid, Anbrazaitis, Svensson-Lundmark, and House present a method for the automatic detection of head movements in audiovisual recordings of read news in Swedish. Head movements are annotated manually, and a machine learning classifier is trained to predict presence or absence of head movement with reference to words based on automatically extracted velocity and acceleration features.

Detection of head movement is also the topic of the paper by Jongejan, Paggio, and Navarretta. In their approach, an SVM classifier learns to classify head movements based on three movement features, i.e. velocity, acceleration, and jerk. The paper also explains that the results of the classifier can be integrated seamlessly with the annotation produced by the ANVIL annotation tool.

Still in the area of automatic recognition, the paper by Stefanov and Beskow describes a method for automatic recognition of isolated Swedish Sign Language signs for the purpose of educational signing-based games. Signs are performed by experienced and inexperienced adult signers and captured with a Kinect sensor. A recognizer based on manual components of sign language is tested on these datasets, with good recognition rates for signer-dependent signing and encouraging ones for signer-independent signing.

A number of papers deal with the use of different non-verbal signals. Among these, the study by Navarretta deals with the relation between audience response and the use of filled pauses and gestural behaviour in humorous speech. The topic is discussed based on statistical analysis of data from two semi-automatically annotated speeches by Barack Obama.

Audience engagement is also investigated in the paper by Curtis, Jones, and Campbell. Their paper looks at correlations between emphasised speech, and increased levels of audience engagement during audio-visual presentations and lectures. In turn, the authors show that emphasised speech corresponds to video sequences characterised by an increase in pitch accompanied by high visual motion.

Jehul, Brône, and Feyaerts study the use of the fillers ‘euh’ and ‘euhm’ in Dutch conversational data, particularly how they co-occur with gaze. Based on this multimodal analysis, they distinguish between a deeper cognitive thinking function, mainly associated with the nasal filler, and a word search function reserved for the pure vocalic filler. Their findings confirm a similar distinction also described in other Germanic languages.

The study by Hoetjes presents an experimental investigation of how informative gestures are when produced in repeated references after negative feedback. The results of the experiment show that, when presented with gestures produced after negative feedback, subjects are more likely to identify correctly the object referred to by the gesture, thus indicating that gestures after negative feedback become more informative.

Hand gestures are also the topic of the paper by Wessel-Tolvig and Paggio. This paper presents the results of two judgment tasks in which the authors investigate how Italian speakers understand complex prepositions in ambiguous motion event constructions. The results show that these constructions are subject to certain semantic constraints, but also that co-speech gestures change the reading of event constructions and thus override default meaning expressed only in speech.

The two papers by Ousterhout both deal with how the semantic processing of emojis can be studied by means of EEG technology. The first of these papers, on BCI effectiveness, establishes the viability of commercial EED equipment to measure the N400 effect, which is known to be an effective way of detecting processing of semantically incongruous signals. The second paper investigates, through survey data and measurements obtained with a commercial EEG, how users understand moving facial emojis occurring in different sentence positions. The results indicate that there is no preference for particular positions of the emojis, and that some of the unusual emojis are ambiguous and are presumably ignored in the processing of the sentence in which they are embedded.

A qualitative methodology is largely followed in the paper by Gerwing, which focuses on the use of gesture, particularly body-oriented gesture, in the area of healthcare. The data for the study come from publically available videos from actual patient-physician encounters. About 150 examples of body-oriented gestures are found and analysed, and it is shown that these gestures have a range of functions related to grounding and cohesion in the interactions.

The paper by Allwood, Ahlsén, Lanzini, and Attaran is also related to the area of healthcare. It compares video health information about overweight and obesity in two different countries – Sweden and Malaysia. Interesting differences both at the level of content and representation styles are described, pointing to possible cultural differences in the rhetorical approach to the topic.

In the study by Vincze and Poggi, videotaped monologues in which doctors and scientists explain their scientific findings are analysed to understand how epistemic stance is expressed through words

and body markers. The authors focus in particular on signals of high certainty and obviousness, which they illustrate and discuss from a semantic and cognitive point of view.

Poggi and Ansani take the reader to the realm of music practice, by analysing gestures of intensity in orchestra and choir conduction. Based on examples from a corpus of concerts and rehearsals, they posit a multimodal lexicon of musical intensity where postures and gestures are categorised along a number of dimensions. These gesture types are shared by different conductors, and form thus a specific lexical area governed by systematic semiotic devices.

A semiotic methodology is also followed by Kreydlin and Khesed, who are interested in how parts of the body are expressed in words and non-verbal expressions. In particular, they discuss how 'hair' is conceptualised and expressed in the Russian language, including the expression of actions involving hair.

In conclusion, this collection of papers provides a snapshot of a field in which research from quite different disciplines can hopefully create synergies and knowledge that can contribute to the further development of multimodal studies, as well as a continuing interest for the MMSYM symposia.

Patrizia Paggio and Costanza Navarretta

Copenhagen, 26 June 2017

Towards classification of head movements in audiovisual recordings of read news

Johan Frid
Lund University
Humanities Laboratory,
Lund University, Sweden
johan.frid@humlab.lu.se

Gilbert Ambrazaitis
Linguistics and Phonetics,
Centre for Languages and Literature,
Lund University, Sweden
gilbert.ambrazaitis@ling.lu.se

Malin Svensson-Lundmark
Linguistics and Phonetics,
Centre for Languages and Literature,
Lund University, Sweden
malin.svensson_lundmark@ling.lu.se

David House
Department of Speech, Music and Hearing, KTH, Sweden
davidh@speech.kth.se

Abstract

In this paper we develop a system for detection of word-related head movements in audiovisual recordings of read news. Our materials consist of Swedish television news broadcasts and comprise audiovisual recordings of five news readers (two female, three male). The corpus was manually labelled for head movement, applying a simplistic annotation scheme consisting of a binary decision about absence/presence of a movement in relation to a word. We use OpenCV for frontal face detection and based on this we calculate velocity and acceleration features. Then we train a machine learning system to predict absence or presence of head movement and achieve an accuracy of 0.892, which is better than the baseline. The system may thus be helpful for head movement labelling.

1 Introduction

This study was conducted in the context of a research project on multimodal, or audiovisual, prosodic prominences and their utilization for information structure coding. In particular, the project investigates how head and eyebrow movements interact with sentence-level pitch accents (also referred to as focal accents in the Swedish prosody research community, cf. Bruce (1977), Bruce & Granström (1993)). A crucial part of the project's aim is to explore possible co-occurrences of the three prominence cues (focal accents, head beats and eyebrow beats). For this purpose, annotations of focal accents, as well as head and eyebrow beats are required. Focal accents are assigned to words: functionally, a focal accent lends prominence to a word, and a word can normally only receive one focal accent. Therefore, the domain of interest in the present context is the word, and for that reason, we have decided to define the word as a domain also for annotations of head and eyebrow movements.

One challenge of such a project lies in the annotation of head and eyebrow movements based on video data, which is commonly achieved by means of manual labelling by human annotators. In order to enable future large-scale investigations of multimodal prominence, we are developing automatic methods for the annotation of movements, in this study strictly focusing on head beats.

To this end, we developed a system for training a classifier to recognise head movements in video data. The purpose of the present study is twofold: 1) to see how well we can classify head movements with an automatic classifier, and 2) to identify labelling-related problems. As motivated above, in the present study we use the domain of the word (rather than, e.g., syllable) because of the relation to information structure.

2 Method

Here we describe our procedures for data collection, annotation, video analysis, feature extraction and machine learning.

2.1 Material

Our materials consist of Swedish television news broadcasts and comprise audiovisual recordings of five news readers (two female, three male) from 144 different sessions. The total duration of the recordings is just over 27 minutes and there are about 4200 spoken words in total. There is always only one person present in the video frame at a given time and he/she almost always faces the camera. Hence, face detection is rather straightforward in this material. The frame rate was 25 fps.

2.2 Annotation

This corpus was manually labelled, applying a simplistic annotation scheme consisting of a binary decision about absence/presence of a movement in relation to a word: to this end, the audio-visual data was first segmented at the word level based on the audio data. Praat (Boersma & Weenink, 2014) was used for this purpose. Each segment was also labelled with the actual word spoken.In total, there were 4208 words. There were also 234 sentence- or phrase-internal pauses. These were also annotated and included in the material as they also may be associated with head movements. In total there were 4442 word units. In the rest of this article, we shall refer to them simply as 'words'.

In the next step, ELAN (Wittenburg et al., 2006) was used to determine for each word if there was head movement or not, where ‘presence’ was defined as an event in which the head rapidly changed its position, roughly within the temporal domain of the word. This was done based on the complete audio-visual display.

Our simple annotation scheme (i.e. assigning annotation to words directly) introduces a problem which results in slight discrepancies: As movements may be realized near the border between two adjacent words, or even span two words, the decision as to which of the words should be annotated for the movement in question is not always obvious.

The material was annotated in three different sets by five annotators. Set 1 consisted of 77 sessions (2554 words) and was annotated by annotator 1, Set 2 consisted of 36 sessions (851 words) and was annotated by annotator 2, and finally, Set 3 had 31 sessions (1037 words) and was annotated (independently) by annotators 3-5. For Set 3, an annotation was counted as such in the event of an agreement between at least two annotators ('majority vote'). Furthermore, for Set 3, the absolute agreement (when all three annotators agreed) was 82.7% and Fleiss' κ (Fleiss, 1971) was 0.69.

As it is possible that annotators behave differently, we will look at each annotator group separately as well as the combination of all three sets. For a more detailed discussion of our definition of beat head movements and our other multi-modal annotations (eyebrow beats and verbal prosodic prominence), see Ambrazaitis et al. (2015).

2.3 Video and head movement analysis

For the video analysis we used the frontal face detection functions in the OpenCV library (Viola & Jones, 2001) to detect areas with faces. This method is similar to Zhang et al. (2007). Each frame in the visual speech corpus is analysed, and this gives us an estimate of the location of the face - and head; they are almost equivalent in this context - as coordinates in the x-y plane, as illustrated in Figure 1.

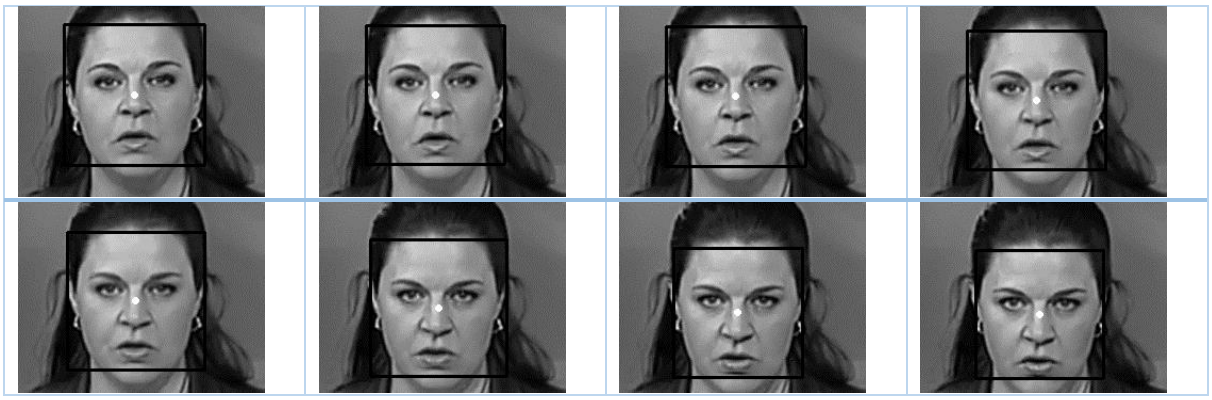


Figure 1. Faces detected in successive frames during a head movement. The black square is the detected face, the white dot (at the center of the square) is the x-y coordinate we use.

The next step is to smooth and calculate velocity and acceleration profiles from the head coordinates. Here we use a method described by Nyström and Holmqvist (2010). We use the Savitzky–Golay (SG) FIR smoothing filter, which makes no strong assumption on the overall shape of the velocity curve and is reported to have a good performance in terms of temporal and spatial information about local maxima and minima (Savitzky & Golay, 1964). Given raw head coordinates this outputs smoothed velocity and acceleration for the x- and y-dimensions separately. Then the total magnitudes of velocity and acceleration are calculated as the Euclidean distance of the x- and y-components. This is shown in Figures 2 and 3, where we also show how we can compare the movement functions with the intervals of our word-related head movement labelling.

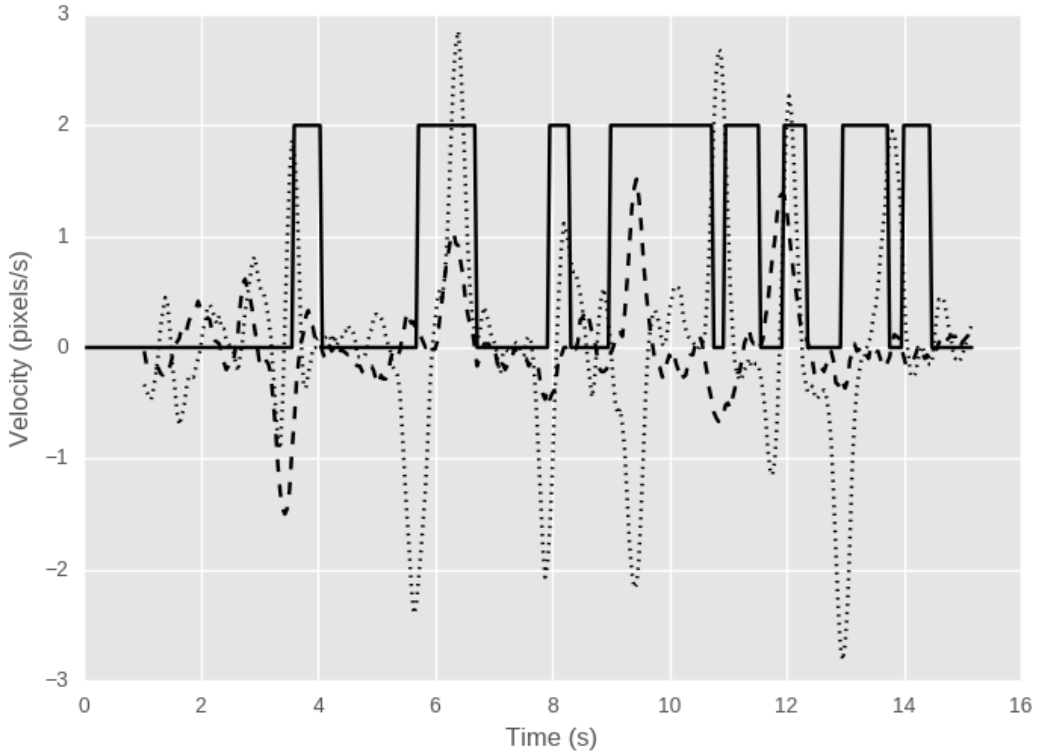


Figure 2. X-velocity (dashed), y-velocity (dotted) and word intervals (solid) as a function of time. The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

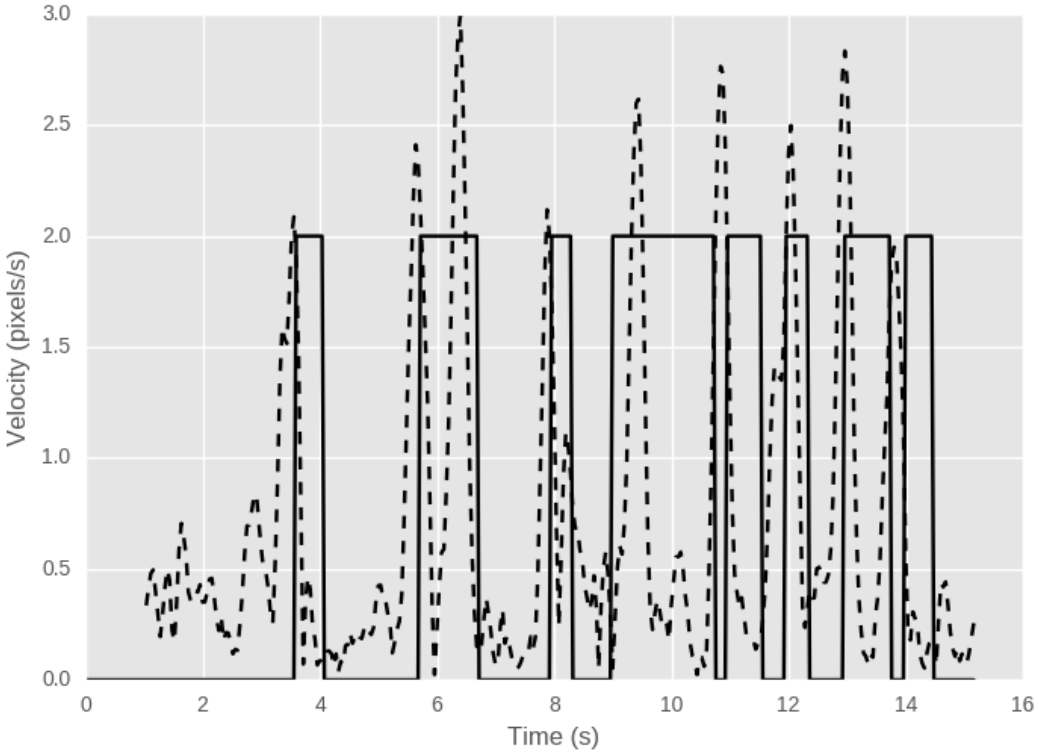


Figure 3.Magnitude of velocity (dashed) and word intervals (solid) as a function of time. The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

2.4 Feature extraction

From each of the six curves (x-velocity, y-velocity, x-acceleration, y-acceleration, magnitude of velocity and magnitude of acceleration) we calculate four features per word: average, max, min and amplitude (max-min). This gives us a total of 24 features.

2.5 Classifier

We then trained a classifier by feeding the features into a machine learning algorithm and training it to predict the outcome movement or no movement. We used Xgboost (Chen & Guerstin 2016), which is a popular method in the machine learning community. It is an ensemble of decision trees, where special care is taken to avoid overfitting. The manually annotated data was used to train and evaluate the classifiers.

3 Results

We ran different experiments, both on all data combined and on subgroups, where the data is split on annotator group or news reader. For the annotator groups, we did not perform any experiments on the individual annotations by annotator 3-5; these are collapsed into 'Annotator 3' in the experiments. Finally, as noted in section 2.2, we know that movements can cross word boundaries, we look at a case where we set the label of neighbours of words annotated with 'movement' to 'movement' (regardless of what they were before). In other words, we let the positively annotated words 'leak' into its neighbours. In this way, we may to some extent capture cases where the movements are crossing word boundaries, but where only one of the words has been labelled 'movement'.

All our experiments are run with xgboost, using 10-fold cross-validation. As evaluation measurements we use accuracy (ACC-XGB), F1 score (F1) and area under ROC curve (AUROC). For comparison, we also calculate the accuracy of a baseline classifier (ACC-BL) that always predicts the majority case. We also include the total number of words (N) and the distribution between movement (N(M)) and no-movement (N(NM)) classes in the tables that follow.

3.1 Combined

Results for all words are presented in Table 1. We note that the xgboost classifier outperforms the baseline classifier.

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Combined	4442	720	3722	0.838	0.892	0.624	0.756

Table 1. Results for all words.

3.2 Subgroups: annotators

Results per annotator are in Table 2. We again note that the machine learning classifier is better than the baseline, and also that the evaluation measurements are both lower than and higher than the 'Combined' case (Table 1).

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Annotator 1	2554	395	2159	0.846	0.885	0.556	0.717
Annotator 2	851	96	755	0.887	0.936	0.666	0.803
Annotator 3	1037	229	808	0.779	0.865	0.667	0.786

Table 2. Results per annotator.

3.3 Subgroups: news readers

We show the results split per news reader in Table 3. The xgboost classifier again gets higher accuracy score than the baseline, and compared with the 'Combined' case (Table 1) we again see the results going in both directions.

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Newsreader 1	904	80	824	0.916	0.932	0.542	0.721
Newsreader 2	981	216	765	0.780	0.857	0.672	0.746
Newsreader 3	508	65	443	0.872	0.920	0.661	0.800
Newsreader 4	1318	238	1080	0.819	0.878	0.618	0.755
Newsreader 5	731	121	610	0.834	0.892	0.647	0.785

Table 3. Results per newsreader.

3.4 Neighbours

Finally, the results for the 'Neighbours' case are shown in Table 4. The xgboost classifier again outperforms the baseline if we compare their accuracy. We also note that the difference is much larger than in the 'Combined' case (Table 1).

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Neighbours	4442	1817	2625	0.59	0.738	0.653	0.718

Table 4. Results for all words, with neighbours changed.

4 Discussion

Overall, our system performs better than the baseline, which we take as an indication that it might be useful for labelling new, unknown data.

As regards the differences between the annotators, if the performance of the system had been better for each individual annotator, this would mean that there would be an annotator-dependent pattern that would disfavour grouping all data together. Since this is not the case, we can use different labellers (or labeller groups).

Similarly, if all the results for individual news readers had been better than the 'Combined' case, then our data would not have any generative power. Since this is not the case, we think our method performs well for the general case where all news readers are combined and would be applicable to other news readers.

The classifier may be helpful for head movement labelling in its own right. Moreover, as mentioned in section 2.2 and shown in Figures 2 and 3, our labelling poses some problems for the classifier: we see that there are cases where the peak of the velocity curve crosses the word label function. This means that the head movement occurs right on a word boundary. This is a problem as one word then

has been labelled as 'movement' and the other as 'no movement', but both may have large velocity/acceleration. Our 'Neighbours' condition is one attempt to deal with that, and we think that the fact that the improvement over the baseline is larger in this condition indicates that it is useful to look at possibilites beyond the word. We intend to pursue other strategies for this in the future.

Another, less problematic, case is that more than one head movement can co-occur with the same word. Our feature extraction deals with that as it is not dependant upon the number of peaks within a word, just the max, the average etc.

5 Conclusion

We have developed a system for the detection of word-related head movements in audiovisual recordings of read news. The task seems feasible; our data seems to have predictive power. The results show no effects from using individual vs groups of labellers. Furthermore, they show that it is possible to generalize over several different news readers. Labelling at word boundaries causes some issues when head movements occur across boundaries.

Acknowledgements

This work was supported by an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2014–2018; grant number 821-2013-2003), a grant from the Marcus and Amalia Wallenberg Foundation (grant number 2012.0103) and also partially funded by the Bank of Sweden Tercentenary Foundation (grant number P12-0634:1). We also thank our two additional annotators Anneliese Kelterer and Otto Ewald.

Reference

Ambrazaitis, G., Svensson Lundmark, M. & House, D. (2015). Multimodal levels of prominence : a preliminary analysis of head and eyebrow movements in Swedish news broadcasts. In Svensson Lundmark, M., Ambrazaitis, G. & van de Weijer, J. (Eds.) Working Papers in General Linguistics and Phonetics (Proceedings from Fonetik 2015) (pp. 11-16), 55. Centre for Languages and Literature, Lund University.

Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bruce, G. 1977. Swedish Word Accents in Sentence Perspective. Travaux de l'institut de linguistique de Lund 12. Malmö: Gleerup.

Bruce, G., B. Granström (1993). Prosodic modelling in Swedish speech synthesis. Speech Communication 13, 63–73.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Fleiss, J. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378-382.

Nystrom, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behavior Research Methods, 42, 188-204. doi:10.3758/BRM.42.1.188

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36, 1627-1639.

Viola, P., & Jones, M. J. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Volume: 1, pp.511–518.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. Proc. of LREC 2006, Fifth International Conference on Language Resources and Evaluation. See also: http://tla.mpi.nl/tools/tla-tools/elan/

Zhang, S., Wu, Z., Meng, H., Cai, L. (2007) Head Movement Synthesis based on Semantic and Prosodic Features for a Chinese Expressive Avatar In: ICASSP 2007, Vol. 4, pp.837-840, 2007.4

Classifying head movements in video-recorded conversations based on movement velocity, acceleration and jerk

Bart Jongejan
University of Copenhagen
bartj@hum.ku.dk

Patrizia Paggio
University of Copenhagen
paggio@hum.ku.dk
University of Malta
patrizia.paggio@um.edu.mt

Costanza Navarretta
University of Copenhagen
costanza@hum.ku.dk

Abstract

This paper is about the automatic annotation of head movements in videos of face-to-face conversations. Manual annotation of gestures is resource consuming, and modelling gesture behaviours in different types of communicative settings requires many types of annotated data. Therefore, developing methods for automatic annotation is crucial. We present an approach where an SVM classifier learns to classify head movements based on measurements of velocity, acceleration, and the third derivative of position with respect to time, *jerk*. Consequently, annotations of head movements are added to new video data. The results of the automatic annotation are evaluated against manual annotations in the same data and show an accuracy of 73.47% with respect to these. The results also show that using *jerk* improves accuracy.

1 Introduction

This paper is about the automatic annotations of head movements in multimodal videos of face-to-face conversations. Head movements are the most frequent gestures in face-to-face communication, where they have numerous functions, most of them related to the management of the interaction (Allwood, 1988), especially feedback giving, also known as backchannelling (Yngve, 1970; Duncan, 1972), as well as turn management (McClave, 2000).

Since head movements are important social and communication signals, their uses in different types of communicative settings and their automatic recognition have been addressed by many researchers the past decades (Heylen et al., 2007; Paggio and Navarretta, 2011; Morency et al., 2007).

Some of this work makes use of human annotators to identify and categorise gestural behaviour, specifically head movements. Manual annotation of gestures, however, is resource consuming, which probably explains why there is still a lack of large annotated multimodal corpora in different languages and communicative settings. Such data are important for modelling human behaviour and implementing natural human-machine interactions. Therefore, developing automatic annotation methods in this area is crucial.

The method for automatic head movement annotation described in this paper is implemented as a plugin to the freely available multimodal annotation tool ANVIL (Kipp, 2004), using OpenCV (Bradski and Koehler, 2008). It builds on earlier work (Jongejan, 2012), where thresholds for velocity and acceleration were used to detect head movements, and extends that work in two important ways: i. by adding jerk to the movement features taken into account; ii. by recasting the problem in machine learning terms.

2 Background

Research aimed at the automatic recognition of head movements, especially nods and shakes, has addressed the issue in two fundamentally different ways either by using data in which the face, or a part of it, has been tracked via various devices, or by working with raw video material. For example, Kapoor and Picard (2001) identify nods and shakes through the position of eye pupils obtained via an infrared camera. The data for this study was collected asking ten participants to answer yes-no questions with nods and shakes. An HMM model trained on this data achieved a prediction accuracy of 75% for nods and 81.02% for shakes. Similarly, Tan and Rong (2003) identify a point between the eyes by means of eye tracking and use this position to recognise nods and shakes. An HMM model trained on data

containing 37 nods and 49 shakes achieved an accuracy of 82% for nods and 89% for shakes. Still in the area of tracking-supported prediction, Wei et al. (2013) use data obtained via Kinect sensors to detect head nods and shakes. They report 86% accuracy.

While the use of tracked data yields relatively good accuracy for the recognition of head movements, this approach requires the use of tracking devices in specific settings and lighting conditions. Therefore, in parallel with this research, other studies have addressed the automatic identification of gestures from raw video material.

One of the techniques used for the automatic recognition of both head and hand gestures from videos is optical motion flow, which makes it possible to identify faces by skin colour segmentation. For example, (Zhao et al., 2012) determine the position of nostrils in videos via optical motion flow and use this position to identify head movements in a corpus of video fragments in which 10 participants were asked to perform repeated nods, shakes, head bows and turns in a predefined order. A boosting algorithm was trained on part of the videos and then tested on the remaining part. The authors report accuracy results of 100% for nods and 84% for bows. These results, however, do not address naturally occurring gestures in conversations. In fact, the limitation of optical motion flow is that it only works well if the videos are recorded in controlled environments since it is very sensible to light and background conditions.

Morency et al. (2005) use a head pose tracker, WATSON, which returns three angular head velocities, and they train an SVM algorithm on frequency-based features of these velocities. Their training data consisted of 10 natural head movement sequences from recorded interactions of humans with an embodied agent, MEL. They also used 11 posed gesture sequences as additional training data. Their system was then tested on 30 video recordings of 9 participants interacting with a robot. The authors report true detection rates of 75% for nods and 84% for shakes for a fixed false positive rate of 0.05. In a later study, Morency et al. (2007) use a Latent-Dynamic Conditional Random Field (LDCRF) model to detect visual gestures in the same data. The accuracy of the new model ranged from 65% to 75% for a false positive rate of 20-30% and outperformed both SVM and HMM models.

Al Moubayed et al. (2009) use OpenCV to detect faces and apply the Lucas-Kanade algorithm to compute velocity as a function of time for identifying smiles from videos. In previous work (Jongejan, 2012) , we apply OpenCV detection of faces and use velocity and acceleration measures, in combination with customisable thresholds, for the automatic annotation of head movements in ANVIL (Kipp, 2004). The obtained annotations were compared with manual annotations and it was found that they correlated well, allowing for the fact that the automatic annotations systematically anticipated movement onset of a few frames compared with the manual annotations. Therefore, the approach was considered promising, in spite of the fact that the algorithm tends to find many small movements compared to the few longer ones identified by the annotators, and notwithstanding the high number of false positives, an issue also raised in Jokinen and Wilcock (2014).

The present work is still based on the use of physical characteristics of the head movements, i.e. velocity and acceleration. However, jerk is added as a third feature. Furthermore, the use of manually defined thresholds to guide movement identification is abandoned in favour of a machine learning approach.

3 Velocity, acceleration and jerk

Three derivatives with respect to time of the position of the face are used in this work as features for the identification of head movements: velocity, acceleration and jerk. Velocity is change of position per unit of time, acceleration is change of velocity per unit of time, and finally jerk is change of acceleration per unit of time. We expect that a sequence of frames for which jerk has a high value in the horizontal or vertical direction will correspond to the most effortful part of the head movement (often called *stroke* (Kendon, 2004), or *apex* (Loehr, 2007)).

Fig. 1 and Fig. 2 both illustrate how a constant, positive jerk can be used to model a nod in the course of almost one second. Fig. 1 depicts the effect of jerk in connection with an idealised nod that starts from rest and that initially is under the influence of a negative (downward directed) acceleration. We see that the downward acceleration causes the head to move down at an increasing rate. As a result of the positive jerk, however, the downward acceleration weakens and turns into an upward, increasing

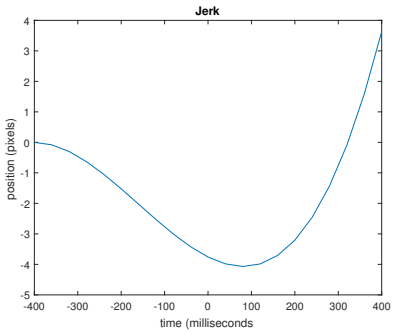


Figure 1: Jerk in an idealised nod. The figure depicts the relative position of the head in a time period of 800 milliseconds. Initially, at -400 ms, the head is at rest in position 0. Due to a negative (downward) acceleration it moves a few pixels down, but the positive, constant (43.74 $pixels/s^3$) jerk changes the acceleration in the positive sense, first weakening it until reaching zero (at -159.2593 ms) and then strengthening it in positive direction, stopping the downward movement at t=81.4815 and then turning it in an upward movement, passing the initial position at t=322.222 ms.

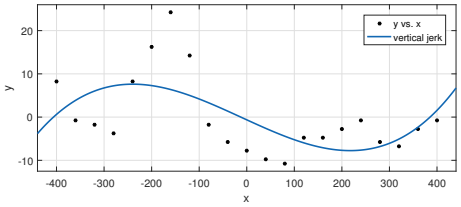


Figure 2: Jerk during a nod computed from data produced by the Anvil Facetracker. The figure depicts the relative vertical position of the head in a time period of about 800 milliseconds, or 21 frames (assuming a frame rate of 25 frames/s). Points on the curve approximate the relative position of the head, the steepness of the curve indicates the head's velocity. The acceleration is downwards in parts where the curve curls downward (left half of the curve) and is upwards in parts where the curve curls upward (right half of the curve). The jerk is positive.

acceleration. After a short delay, the upward acceleration first stops and then reverses the downward movement. As a whole, the curve seems a good model of the type of movement we understand as a nod, where the head, after having bounced down, quickly accelerates upwards.

In reality, a head movement rarely starts from total rest. Fig. 2 illustrates a typical sequence of real data points and the jerk that we compute from them. The computation is based on the measurements of the vertical position of a head during a time window of 21 frames, or about 800 ms. Although the person's head movement is smooth, the data points are jumping and they look in some places like outliers, which is partly due to technical imperfections during filming and partly due to algorithmic inaccuracies.

As in Fig.1, the jerk is positive and assumed to be constant, but here the movement does not start from rest. After a short upwards trajectory, the movement proceeds downwards and then up just as in Fig. 1.

4 Data, training and test setup

The data used for this work is a subset of the videos recorded and annotated in the Nordic NOMCO corpus (Paggio et al., 2010), and in particular the Danish part of the corpus, a collection consisting of twelve videos in which pairs of speakers who never met before (six males and six females) are seen chatting freely for about five minutes. Each speaker took part in two different conversations, one with a male and one with a female. The speakers are standing in front of each other on a carpet, which delimits the space between them. The conversations were recorded in a studio using three different cameras and two cardioid microphones. The data were subsequently annotated with many different annotation layers (Paggio and Navarretta, 2016), including temporal segments corresponding to different types of head

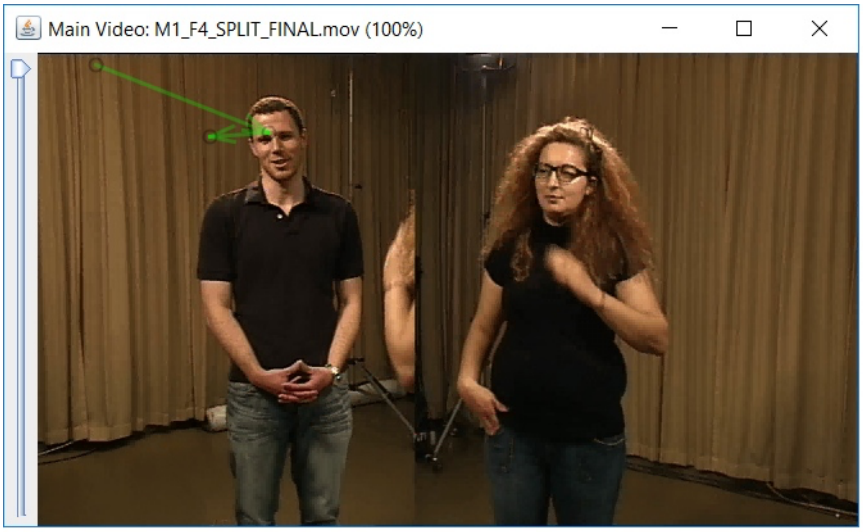


Figure 3: A frame in the midst of an HeadOther movement. The arrow indicates the strength and direction of the jerk movement feature

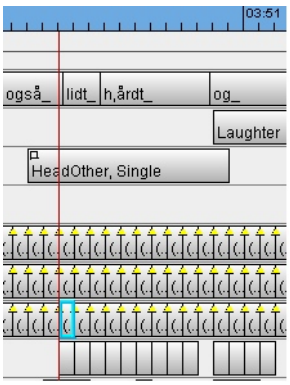


Figure 4: Area in Annotation window showing from top to bottom the manual annotation for the HeadOther movement in the previous figure, three tracks of frame-by-frame annotations for velocity, acceleration and jerk, and the annotation predicted by the SVM model. It can be seen that the Laughter annotation also coincides with a predicted head movement.

movement. The types distinguished in the coding scheme (Allwood et al., 2007) and annotated in the corpus, are *Nod, Up-nod, Shake, Turn, Tilt, Headbackwards, Headforwards, Waggle*, and *HeadOther*, and each head movement has also been coded with one the two features *Single, Repeated*. The inter-annotator agreement obtained for these annotation attributes is on average a κ-score of 0.61.

For this work, two videos sharing one of the speakers were selected at random, and only the head movements performed by this one shared speaker were considered. In both videos, OpenCV was used to analyse each frame for the x and y (horizontal and vertical) coordinates of the speaker. This was done interactively using the ANVIL tool, through a dedicated plugin developed for this purpose. The researcher can follow the process through a visual rendering of how the observable characteristics are detected (Fig. 3), and also through the annotation board (Fig. 4). The coordinates were buffered so that velocity, acceleration and jerk could be computed with reasonable accuracy. For velocity we include the previous three and next three frames in the computation, giving a total of seven frames. For velocity and jerk we need even more frames to reduce the effect of noise in the data to an acceptable level, 14 and 21 frames, respectively.

After having added this frame-wise annotation, one of the video was used as training data, and the

margin	characteristics	accuracy	baseline	precision	recall	F-score
0	VAJ	68.81	64.15	71.48	21.66	33.25
0	VA	67.62	64.15	67.56	18.66	29.24
1	VAJ	69.15	64.15	71.82	22.98	34.82
1	VA	67.78	64.15	67.12	19.88	30.67
2	VAJ	69.52	64.15	72.22	24.34	36.41
2	VA	68.06	64.15	67.53	21.04	32.08
3	VAJ	69.75	64.15	71.70	25.85	38.00
3	VA	68.48	64.15	68.11	22.72	34.07
15	VAJ	71.59	64.15	69.11	37.55	48.66
15	VA	69.39	64.15	64.56	32.43	43.17
16	VAJ	71.64	64.15	68.68	38.45	49.30
16	VA	69.53	64.15	64.36	33.67	44.21
17	VAJ	71.50	64.15	67.87	38.94	49.49
17	VA	69.77	64.15	64.64	34.65	45.12
18	VAJ	**71.65**	64.15	67.38	40.59	**50.66**
18	VA	69.99	64.15	64.67	35.92	46.19
19	VAJ	71.24	64.15	65.74	41.31	50.74
19	VA	69.76	64.15	63.59	36.65	46.50

Table 1: Several statistics obtained from training an SVM model on OpenCV output from one video and testing the SVM model on OpenCV output from another video with the same person standing on the same spot and generally facing in the same direction. Each video is about 8 minutes long and has a rate of 25 frames per second. VAJ stands for velocity, acceleration, and jerk. VA stands for velocity, and acceleration.

other was set aside as test data.

To complete the preparation of the training data, each frame was supplemented with the feature '1' if it was included in a head movement in the manually annotated file, and with the feature '0' otherwise.

A first inspection of the results of this initial frame-wise annotation revealed that in several cases, OpenCV detected sequences of movement interrupted by empty frames, where the manual annotation consisted of longer spans of uninterrupted movement. Therefore, we experimented with allowing empty spans of varying length to be considered part of the movement annotation in the subsequent machine learning experiments. Such a span of one or several frames is called *margin* in what follows.

5 Results and discussion

SVM classifiers were trained with a range of different margin values, and using all three movement characteristics together (VAJ), only velocity and acceleration (VA), as well as each individual characteristic alone (V, A, and J). The best performing classifiers were those using all three characteristics, followed by those using two, therefore only results obtained using these two alternatives will be discussed here.

Table 1 compares results obtained by using velocity, acceleration and jerk data with results obtained by only using velocity and acceleration data. It does so for a selection of values for the interpolation margin. If there are no more than <margin> negative frames between two positive frames, the negative frames are converted to positive frames. The first column mentions the size of the margin, the second the characteristics that were used to train and test the SVM model, where V=velocity, A=Acceleration, and J=Jerk. Accuracy is the proportion of correct labels (true positives and true negatives) assigned by the model. The baseline is the accuracy obtained by always assigning the non-movement label. Precision is the number of true positives over the total movement labels assigned (true and false positives). Recall is the number of true positives over the total movement labels that should have been assigned (true positives and false negatives). The F-score corresponds to F_1.

The best results both in terms of F-score (50.66) and accuracy (71.65) are obtained choosing VAJ,

with a relatively high value of the margin, 18 frames, or 0.72 seconds. Note that OpenCV was not able to detect a face in all frames. This results in about 3 percent lower accuracy overall.

Head movement	# frames	accuracy (%)	min (%)	max (%)
(no movement)	6190	95.17	0.00	100.00
HeadForward	239	8.37	0.00	38.10
Waggle	30	50.00	30.77	64.71
HeadOther	435	34.83	0.00	85.00
Nod	244	28.69	0.00	92.86
Jerk	85	22.35	0.00	60.00
Tilt	696	21.98	0.00	87.50
SideTurn	1134	21.08	0.00	88.89
Shake	319	15.36	0.00	43.90
HeadBackward	278	11.87	0.00	53.33

Table 2: Frame-wise detection of movement in different head movement types. Data obtained using an SVM model based on all of velocity, acceleration and jerk, with no postprocessing (margin = 0). "Accuracy" is the overall probability that a frame is correctly recognised as part of a movement or a non-movement, computed as the ratio between the number of recognised frames and the number of all frames in the ensemble of all movements of a given type. "Min" and "max" are the minimum and maximum ratios found in the ensemble of all movements of a given type.

Table 2 illustrates how the accuracy of the SVM method in detecting presence or absence of movement greatly varies over different occurrences of head movements and also over different types of head movements.

The variation in frame-wise accuracy of movement detection in different movement types can be explained in light of a number of observations. First of all, the data used in this study are not of the best technical quality: light conditions are not optimal, the angle from which the subjects are recorded is not completely frontal, and since the output of two different cameras is combined in one video, the two shootings sometimes interfere with each other, as can clearly be seen in Fig. 3. However, in general naturally-occurring multimodal data cannot be expected to conform to high recording standards, and it is therefore a useful exercise to test automatic annotation tools against data of sub-optimal quality. Another issue that not doubt contributed to the lack of agreement between manual and automatic annotation is the fact that only communicative head movements were annotated in the NOMCO corpus. Given this, it is reasonable to expect that the classifier, which works on *any* kind of observable head movement, will detect movement which the annotators decided not to code. Furthermore, when analysing the false positives produced by the classifier we found that over 65% of these annotations fall temporally together with manual annotations of body movements in which also the head changes position as a consequence. These movements were not coded in the manual annotation since they are not independent movements of the head.

Finally, it is important to note that the evaluation presented above is based on frame-wise accuracy, since the developed model is based on frame-wise training. As a result, the accuracy figures reported above cannot be compared to those reported for studies such as Morency et al. (2007), which report accuracy of recognition of whole movements rather than individual frames. Therefore, in Table 3 we report the number of head movements identified by the system with at least some degree of overlap for different movement types. In the table we disregard cases in which the system recognised two or more movements and the manual annotator only coded one as either single or repeated movement. The overall accuracy of 70% is in line with what reported in Morency et al. (2007).

6 Conclusions

We have presented an approach to automatic classification of head movements in raw video data based on the detection of three observable movement characteristics via OpenCV, and the subsequent development

Head movement	# manual	# automatic	% accuracy
Waggle	2	2	100
HeadOther	27	21	78
Tilt	24	15	63
Up-nod	8	4	50
Nod	12	6	50
SideTurn	44	32	73
Shake	14	6	43
HeadBackward	11	6	64
HeadForward	12	8	58
All movements	154	100	**70**

Table 3: Number of head movements identified by the system

of SVM classifiers trained on the head movements of one speaker. The best performing classifier could recognise head movements by the same speaker in unseen video data with an overall accuracy of 73.47%. This accuracy, however, varies considerably for different occurrences and types of head movements.

The best accuracy was obtained using three movement characteristics (velocity, acceleration or jerk), a result which confirms our initial intuition that jerk is a useful feature for the detection of head movements. We have also seen that turning negative predictions into positive ones if a negative prediction has a left and right positive neighbour that are no more than two frames apart, increases accuracy.

Finally, we have demonstrated that the results of the classifiers can be integrated seamlessly with the annotation produced by the ANVIL annotation tool.

Future work will focus on expanding the training material with data from different speakers, experimenting with a more fine-grained classification of movement into horizontal and vertical, as well as considering the distinction between single vs. repeated movements.

References

Sames Al Moubayed, Malek Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, J-C Martin, Stanislav Ondas, Catherine Pelachaud, Jérôme Urbain, and Mehmet Yilmaz. 2009. Generating robot/agent backchannels during a storytelling experiment. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference*, pages 3749–3754. IEEE.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Jean-Claude Martin, Patrizia Paggio, Peter Kuehnlein, Rainer Stiefelhagen, and Fabio Pianesi, 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.

Jens Allwood. 1988. The Structure of Dialog. In Martin M. Taylor, Franoise Neél, and Don G. Bouwhuis, editors, *Structure of Multimodal Dialog II*, pages 3–24. John Benjamins, Amsterdam.

G. Bradski and A. Koehler. 2008. *Learning OpenCV: Computer Vision with the OpenCV Linbrary*. O'Reilly.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud. 2007. Searching for prototypical facial feedback signals. In *Proceedings of 7th International Conference on Intelligent Virtual Agents*, pages 147–153.

Kristiina Jokinen and Graham Wilcock. 2014. Automatic and manual annotations in first encounter dialogues. In *Human Language Technologies - The Baltic Perspective: Proceedings of the 6th International Conference Baltic HLT 2014*, volume 268 of *Frontiers in Artificial Intelligence and Applications*, pages 175–178.

Bart Jongejan. 2012. Automatic annotation of head velocity and acceleration in anvil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 201–208. European Language Resources Distribution Agency.

Ashish Kapoor and Rosalind W. Picard. 2001. A real-time head nod and shake detector. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, pages 1–5, New York, NY, USA. ACM.

Adam Kendon. 2004. *Gesture*. Cambridge University Press.

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

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

Evelyn McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. 2005. Contextual recognition of head gestures. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

P. Paggio and C. Navarretta. 2011. Head movements, facial expressions and feedback in Danish first encounters interactions: A culture-specific analysis. In Constantine Stephanidis, editor, *Universal Access in Human-Computer Interaction - Users Diversity. 6th International Conference. UAHCI 2011, Held as Part of HCI International 2011*, number 6766 in LNCS, pages 583–690, Orlando Florida. Springer Verlag.

Patrizia Paggio and Costanza Navarretta. 2016. The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, pages 1–32.

Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén, Kristiina Jokinen, and Costanza Navarretta. 2010. 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).

W. Tan and G. Rong. 2003. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461–466.

Haolin Wei, Patricia Scanlon, Yingbo Li, David S Monaghan, and Noel E O'Connor. 2013. Real-time head nod and shake detection for continuous human affect recognition. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–578.

Z. Zhao, Y. Wang, and S. Fu. 2012. Head movement recognition based on the Lucas-Kanade algorithm. In *Computer Science Service System (CSSS), 2012 International Conference on*, pages 2303–2306, Aug.

A Real-time Gesture Recognition System for Isolated Swedish Sign Language Signs

Kalin Stefanov
KTH Royal Institute of Technology
TMH Speech, Music and Hearing
Stockholm, Sweden
kalins@kth.se

Jonas Beskow
KTH Royal Institute of Technology
TMH Speech, Music and Hearing
Stockholm, Sweden
beskow@kth.se

Abstract

This paper describes a method for automatic recognition of isolated Swedish Sign Language signs for the purpose of educational signing-based games. Two datasets consisting of 51 signs have been recorded from a total of 7 (experienced) and 10 (inexperienced) adult signers. The signers performed all of the signs 5 times and were captured with a RGB-D (Kinect) sensor, via a purpose-built recording application. A recognizer based on manual components of sign language is presented and tested on the collected datasets. Signer-dependent recognition rate is 95.3% for the most consistent signer. Signer-independent recognition rate is on average 57.9% for the experienced signers and 68.9% for the inexperienced.

1 Introduction

Sign language and different forms of sign-based communication is important to large groups in society. In addition to members of the deaf community, that often have sign language as their first language, there is a large group of people who use verbal communication but rely on signing as a complement. A child born with hearing impairment or some form of communication disability such as developmental disorder, language disorder, cerebral palsy or autism, frequently have the need for this type of communication known as *key word signing*. Key word signing systems borrow individual signs from sign languages to support and enforce the verbal communication. As such, these communication support schemes do away with the grammatical constructs in sign languages and keep only parts of the vocabulary.

While many deaf children have sign language as their first language and are able to acquire it in a natural way from the environment, children that need signs for other reasons do not have the same rights and opportunities to be introduced to signs and signing. We aim at creating a learning environment where children can learn signs in a game-like setting. An on-screen avatar presents the signs and gives the child certain tasks to accomplish, and in doing so the child gets to practice the signs. The system is thus required to interpret the signs produced by the child and distinguish them from other signs, and indicate whether or not it is the right one and if it was properly carried out.

2 Related Work

Sign languages are as complex as spoken languages. There are thousands of signs in each sign language differing from each other by minor changes in the shape of the hands, motion profile, and position. Signing consists of either manual components that are gestures involving the hands, where hand shape and motion convey the meaning or finger spelling, used to spell out words. Non-manual components, like facial expressions and body posture can also provide information during signing. Sign language recognition (SLR) inherits some of the difficulties of speech recognition. Co-articulation between signs, meaning that a sign will be modified by the signs on either side of it and large differences between signers - signer-specific styles (pronunciation in speech), both contribute to increased variation in the signing.

This paper presents a gesture recognition method that attempts to model and recognize manual components of sign language. Therefore, the work cited in this section is restricted to the specific case of tracking-based manual components extraction and modeling, and isolated word recognition using word-level classifier, where hand shape is not explicitly modeled. A comprehensive review of the research on SLR and the main challenges is provided in (Cooper et al., 2011). Manual components of sign language are in general hand shape/orientation and movement trajectories which are similar to gestures. A comprehensive survey on gesture recognition (GR) was performed in (Mitra and Acharya, 2007) and in (Rautaray and Agrawal, 2015).

Data collection is an important step in building a SLR system. Early systems used data gloves and accelerometers to capture the hands' position, orientation and velocity. These were measured by using sensors such as Polhemus tracker (Waldron and Kim, 1995) and DataGlove (Kadous, 1996), (Vogler and Metaxas, 1997). These techniques were capable of producing very accurate measurements with the cost of being intrusive and expensive. These are the main reasons for vision-based systems to become more popular. Vision-based systems can employ one or more cameras or other non-intrusive sensor (e.g. monocular (Zieren and Kraiss, 2004), stereo (Hong et al., 2007), orthogonal (Starner and Pentland, 1995), depth-sensor, such as the Kinect (Zafrulla et al., 2011)). In (Segen and Kumar, 1999) the researchers used a camera and light source to compute depth, and (Feris et al., 2004) used light sources and multi-view geometry to construct a depth image. In (Starner et al., 1998) the authors used a front view camera paired with head mounted camera. Depth can also be inferred using stereo cameras (Munoz-Salinas et al., 2008), or by using side/vertical mounted cameras as with (Vogler and Metaxas, 1998) or (ASL, 2006). There are several projects which are creating sign language datasets - in Germany, the DGS-Korpus dictionary project (DGS, 2010), in the UK, the BSL Corpus Project (BSL, 2010) and in Sweden the SSL Corpus Project (SSL, 2009). Finally, Dicta-Sign (DICTA, 2012) and SIGNSPEAK (SIGNSPEAK, 2012) are European Community's projects aiming at recognition, generation and modeling of sign language.

Hand tracking is another important part of a SLR system. Tracking the hands in sign language conversation is a difficult task since the hands move very fast and are often subject to motion blur. Furthermore, hands are highly deformable and they occlude each other and the face, making skin color based approaches complex. In early work, the hand segmentation task was simplified by colored gloves. Usually these gloves were single colored (Kadir et al., 2004). More natural and realistic approach is without gloves, where the most common detection approach uses a skin color model (Imagawa et al., 1998) and (Awad et al., 2006). Often this task is simplified by restricting the background to a specific color (Huang and Huang, 1998) or keeping it static (Starner and Pentland, 1995). In (Zieren and Kraiss, 2005) the authors explicitly modeled the background. Depth can be used to simplify the problem as in (Hong et al., 2007) and (Grzeszcuk et al., 2000). In (Fujimura and Liu, 2006) and (Hadfield and Bowden, 2012) the hands were segmented under the assumption that they are the closest objects to the camera. Recent work on multi-person pose estimation (Cao et al., 2017) illustrates a system which is capable of tracking the hands in dynamic and cluttered environments.

Early work on SLR applied Artificial Neural Networks (ANN) for modeling. The idea of one of the first papers on SLR (Murakami and Taguchi, 1991) was to train an ANN given the features from a DataGlove and recognize isolated signs. In (Kim et al., 1996) the researchers used DataGloves and Fuzzy Min Max ANN to recognize 25 isolated gestures. The work in (Waldron and Kim, 1995) presented an isolated SLR system using ANN and (Huang and Huang, 1998) presented an isolated SLR system using a Hopfield Neural Network. (Yang et al., 2002) used motion trajectories within a Time Delay Neural Network (TDNN) to recognize American Sign Language. Hidden Markov Models (HMM) (Rabiner, 1989) and (Yamato et al., 1992) are a modeling technique well suited to the problem of SLR. The work by (Starner et al., 1998) demonstrated that HMM are a strong technique for recognizing sign language and (Grobel and Assan, 1997) presented a HMM based isolated sign recognition system. (Vogler and Metaxas, 1997) showed that word-level HMM are SLR suitable. In their following work, (Vogler and Metaxas, 1999) they demonstrated that Parallel HMM (PaHMM) are superior to regular HMM, Factorial HMM and Coupled HMM for recognition of sign language.

The above mentioned systems represent important steps towards general sign language recognition.

However for the specific requirements of our application (Swedish Sign Language, real-time, modular, i.e. easy to integrate into sign-based games and extensible) none of the above systems is applicable. In the following we describe our own implementation which meets these requirements.

3 Method

Dynamic gesture recognition problem involves the use of techniques such as time-compressing templates, dynamic time warping, HMM, and TDNN. A time-domain process demonstrates a Markov property if the conditional probability density of the current event, given all present and past events, depends only on the jth most recent event. If the current event depends solely on the most recent past event, then the process is termed a first order Markov process. This assumption is reasonable to make, when considering the positions of the hands of a person through time.

An HMM is a double stochastic process governed by:

- an underlying Markov chain with a finite number of states,
- a set of random functions, each associated with one state.

In discrete time instants, the process is in one of the states and generates an observation symbol according to random function corresponding to that state. Each transition between the states has a pair of probabilities, defined as follows:

- transition probability, which provides the probability for undergoing a transition,
- output probability, which defines the conditional probability of emitting an output symbol from a finite alphabet when the process is in a certain state.

HMM have been found to efficiently model spatio-temporal information in a natural way. The model is termed "hidden" because all that can be seen is a sequence of observations. An HMM is expressed as $\lambda = (A, B, \Pi)$, where A is state transition probability, B is observation symbol probability and Π is initial state probability. For a classification problem, the goal is to classify the unknown class of an observation sequence O into one of C classes. If we denote the C models by λ_c, $1 \leq c \leq C$, then an observation sequence is classified to class c^*, where $c^* = \text{argmax}_{c \in C} P(O|\lambda_c)$. The generalized topology of an HMM is a fully connected structure, known as an *ergodic* model, where any state can be reached from any other state. When employed in dynamic gesture recognition, the state index transits only from left to right with time, as depicted in Figure 1. Here the state transition probabilities $a_{ij} = 0$ if $j < i$, and $\sum_{j=1}^{N} a_{ij} = 1$.

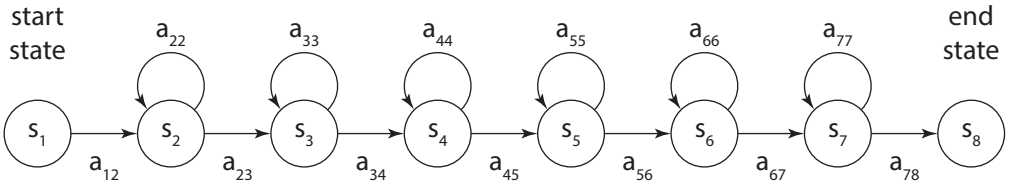


Figure 1: 8-state left-to-right HMM for gesture recognition.

The first step in the method is to train models for all 51 signs. A data recording tool was built to capture participants performing different signs in front of a Kinect sensor. The application and the collected data are described in Section 4. The feature extraction procedure (hands position) depends solely on the skeleton tracking algorithm implemented on the Kinect sensor. The results reported in this work are based on implementation that relies only on the spatial information provided from the Kinect skeleton tracking. We are working on combining skin color-based tracker and the skeleton tracker to achieve better tracking accuracy, as it is obvious that the skeleton tracker commits many errors for certain types of spatial configurations (e.g. the hands are close to each other). Furthermore, as mentioned previously, we are not modeling the shape of the hands explicitly. Let's consider the extraction of the right wrist position

feature. Visualization of the skeleton and the joints used during feature extraction is shown in Figure 2. The 3-dimensional position of the RW is calculated in user-centered coordinate space (the origin of the space coincides with the location of the signer's hip). Then the distance between the shoulders (RS and LS) is used to normalize the feature in **X**. The distance between the neck joint (C0) and the spine joint (C1) is used to normalize the feature in the **Y** dimension. The **Z** dimension of the feature is not normalized. This procedure is done for all six joints under consideration - RH, LH, RW, LW, RE, and LE. Additionally, the first derivative of each feature is calculated in order to compensate for the assumption of observation independence in the state to output probability models of the HMM. In this way a 36-dimensional feature vector is constructed and used as an observation symbol at each time step. The Baum-Welch algorithm (Rabiner, 1989) and (Rabiner and Juang, 1993) is used for training an 8-state sign-level HMM and finding the best model parameters for each sign (in the current implementation we model all signs with the same 8-state HMM topology, which might not be optimal and will be investigated further).

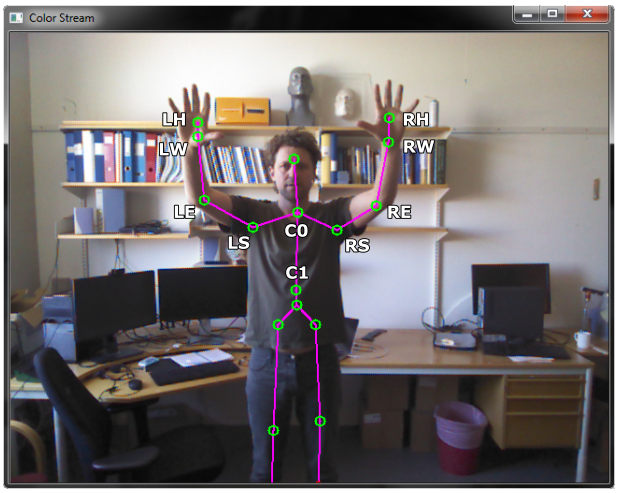


Figure 2: Visualization of the skeleton and the joints used for features extraction.

In this work the observation symbols of the HMM are the extracted feature vectors at each time step. Each class is represented by the model trained for each sign. If we get an unknown observation sequence, we have to calculate $P(O|\lambda_c)$ for all c and classify the sequence into the class which yields the maximum probability. The Viterbi algorithm (Rabiner, 1989) and (Rabiner and Juang, 1993) is applied for finding the most likely sequence of hidden states that results in the observed sequence.

4 Dataset

The size of the recorded vocabulary is 51 signs from the Swedish Sign Language (SSL). The vocabulary is composed of four subsets - objects, colors, animals, and attributes. Table 1 summarizes the vocabulary used in the current implementation.

Objects (17)	**Colors (10)**	**Animals (13)**	**Attributes (11)**
car, tractor, boat, book, ball, doll, computer, flashlight, guitar, watch, necklace, hat, key, patch, scissors, frame, drums	blue, brown, green, yellow, purple, orange, red, pink, black, white	moose, monkey, bear, dolphin, elephant, horse, camel, rabbit, cat, crocodile, lion, mouse, tiger	angry, happy, sad, small, dotted, striped, checkered, narrow, large, thick, tired

Table 1: Vocabulary.

The participants are recorded performing all 51 signs in one session. In total 5 sessions are recorded, resulting in 5 instances of each sign per participant. The recording environment is not explicitly controlled, the only requirement is that the upper body of the participant falls in the field of view of the

Kinect. Some signs involve movements that are difficult for accurate tracking with the skeleton tracking algorithm. Although this introduces errors in the dataset, we kept these instances assuming that similar errors will be committed by the algorithm during real-time usage.

4.1 Data Collection Tool

The system uses the Kinect sensor as input device. For the purpose of recording we created a data collection tool that prompts the participant to perform a certain sign and captures the color, depth and skeleton streams produced by the Kinect. The process of recording a sign follows a certain path. First the meaning of the sign is displayed as text and a video that demonstrates the sign is played, with an option of replaying the video. Once the participant is comfortable enough, the recording of the sign is started. All participants were instructed to start at rest position (both hands relaxed around the body), then, perform the sign and go back to rest position. If the sign is performed correctly the recording moves to the next one, otherwise, the participant can perform the sign until he/she is satisfied.

4.2 Experienced Signers

This part of the dataset is composed of 7 participants (5 sessions each) performing the set of 51 signs. The participants are experts in SSL (sign language researchers/teachers), where six of them are deaf. They were asked to perform the signs in the way they would normally sign. We collected total of 1785 sign instances and used all data in the experiments.

4.3 Inexperienced Signers

This part of the dataset is composed of 10 participants (5 sessions each) performing the set of 51 signs. The participants had no prior experience in SSL. They were instructed on how to perform the signs in two ways - by a video clip played back in the data recording application, and by live demonstration by the person operating the recording session. We collected total of 2550 sign instances, and used all data in the experiments.

5 Experiments

The experiments are done on the dataset described in Section 4. The results are based only on spatial features extracted from the collected skeleton data, Section 3. The conducted experiments are divided into two groups - signer-dependent and signer-independent.

5.1 Signer-dependent

In the signer-dependent experiment we test the recognition rate of the HMM only for separate signers. As described previously, we collected 5 instances of each sign for 7 experienced signers and 5 instances of each sign for 10 inexperienced signers. Table 2 summarizes the results for the experienced signers and Table 3 summarizes the results for the inexperienced signers. All results are based on leave-one-out cross-validation procedure, where the models were trained on 4 instances of each sign for each signer and tested on the 5th instance of the sign performed by that signer.

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	92	96	97.6	99.2	99.2
2	85.5	91.8	93.3	94.9	96.5
3	84.4	92	95.6	96	96.4
4	**95.3**	97.6	98.8	99.6	99.6
5	88	93.2	95.2	96	96.4
6	87.8	94.1	96.5	98.4	98.4
7	**80**	87.5	92.9	95.3	96.5
μ	87.6	93.2	95.7	97.1	97.6
σ	5	3.3	2.2	1.9	1.4

Table 2: Signer-dependent results (7 experienced signers).

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	84.3	92.6	96.9	96.9	97.6
2	92.2	96.1	97.3	98.8	99.6
3	**75.7**	83.9	89.4	92.5	93.7
4	94.9	99.2	99.6	100	100
5	94.5	98.4	99.2	99.2	99.2
6	93.3	96.7	97.6	98	98.8
7	89.4	96.5	97.2	98	98.4
8	**95.3**	98	99.2	100	100
9	94.1	96.5	98.4	98.4	98.8
10	89.8	96.1	97.6	98.4	98.8
μ	90.3	95.4	97.2	98.1	98.5
σ	6.2	4.4	2.9	2.3	1.8

Table 3: Signer-dependent results (10 inexperienced signers).

5.2 Signer-independent

In the signer-independent experiment we test the recognition rate of the HMM between signers. The results shown in Table 4 are based on the data from the experienced signers. We employ leave-one-out cross-validation procedure, where the HMM are trained on 30 instances of each sign (6 different signers) and tested on the 5 instances of the sign performed by the 7th signer. Figure 3 illustrates the confusion matrix for the experienced signers. The matrix is composed of the results for all 7 signers, where the shades of gray of each cell represent the recognition rate for the particular sign. The maximum point in the matrix is 1 which shows that there are signs that are fully recognizable in the signer-independent case, given the simple spatial features. On the other hand, in the figure we can also see that signs which are similar in terms of trajectories are confused (e.g. *drums* - sign #50 and *car* - sign #5).

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	62	80.8	87.6	94.8	98
2	53.3	68.6	77.3	82.3	87.1
3	**48**	65.2	72.6	78.4	81.2
4	**65.5**	82	86.7	90.6	93.7
5	62.6	75.6	79.2	81.4	84.8
6	59.2	69.8	74.9	78.8	83.1
7	54.9	69	78	83.5	88.6
μ	57.9	73	79.5	84.5	88.1
σ	6.1	6.5	5.7	6.2	6

Table 4: Signer-independent results (7 experienced signers).

The results shown in Table 5 are based on the data from the inexperienced signers. We again employ leave-one-out cross-validation procedure, where the HMM are trained on 45 instances of each sign (9 different signers) and tested on the 5 instances of the sign performed by the 10th signer. Figure 4 illustrates the confusion matrix for the inexperienced signers. In this test none of the signs was fully recognizable between signers (maximum of 0.98). Nevertheless, the overall performance compared to the experienced signers increased with 11%.

6 Conclusion and Future Work

As expected, the performance in the signer-independent experiment is significantly lower than in the signer-dependent - 57.9% and 68.9% compared to 87.6% and 90.3% when averaged over all signers. These accuracy rates are however for the full set of 51 signs. In our current application, there is no situation where the recognizer needs to pick one sign from the full set, instead it is always the case of one out of a small number (e.g. choose one out of five animals). For this type of limited recognition tasks, accuracy will increase drastically. Furthermore, we can control the mix of signs in the games, meaning that we can make sure that signs which are confused by the recognizer never appear together - the performance of the signer-independent recognizer will not be a limiting factor in the game. This means that we can have high confidence in the recognitions and the children will be able to advance in the game only after they have learned the correct way to carry out each sign.

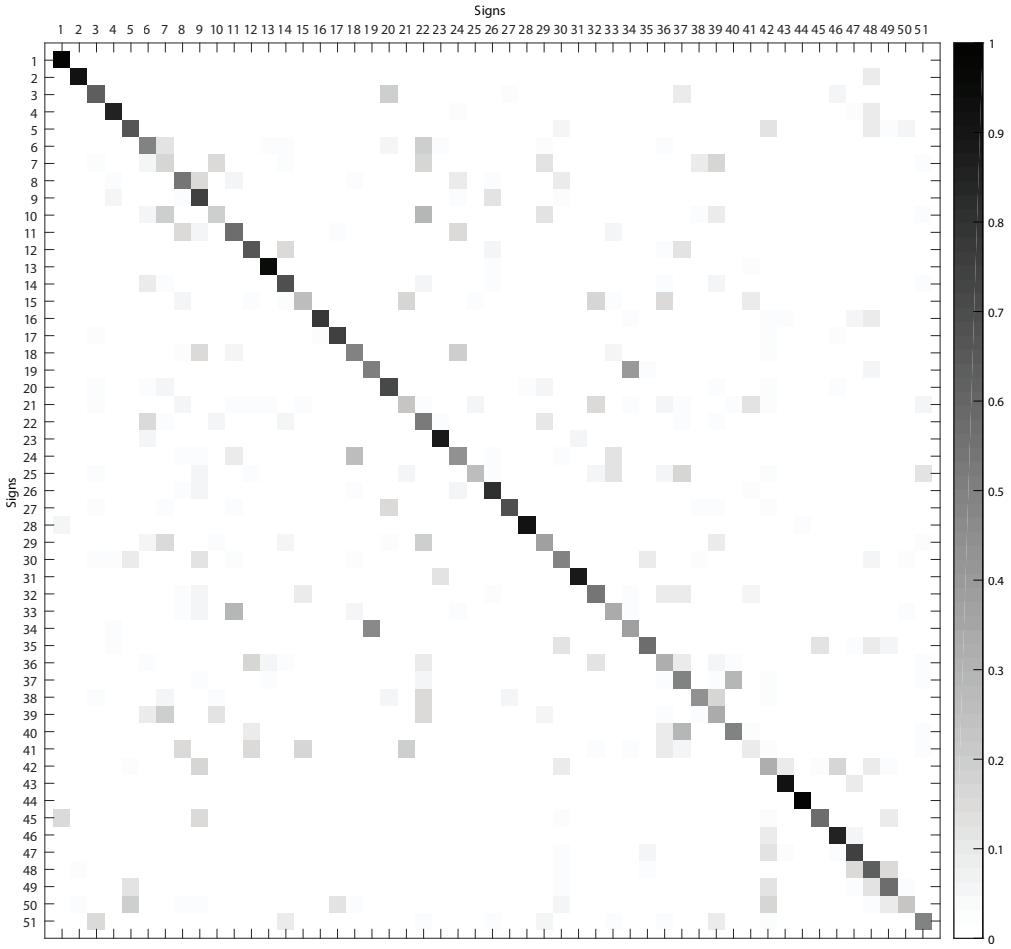


Figure 3: Confusion matrix for all experienced signers.

Participant	1-BEST (%)	2-BEST (%)	3-BEST (%)	4-BEST (%)	5-BEST (%)
1	58.8	73.7	85.9	89.4	92.5
2	72.2	84.7	89.8	91.8	94.1
3	**57.3**	73.7	81.2	88	91.6
4	69.4	80.8	87.4	91	92.5
5	75.3	91	93.7	94.9	97.3
6	**76.1**	85.9	92.5	97.6	98.4
7	73.7	83.1	90.6	91.8	94.9
8	65.5	79.6	84.3	87.4	89.8
9	**76.1**	87.4	91.8	94.5	97.3
10	64.7	79.2	88.2	90.6	94.9
μ	68.9	81.9	88.5	91.7	94.3
σ	7	5.6	3.9	3.2	2.8

Table 5: Signer-independent results (10 inexperienced signers).

An interesting observation is the fact that there is a significant difference in the recognition rate when comparing the experienced and inexperienced signers. One way to explain this is that experienced signers are more casual than their inexperienced counterpart. A similar phenomenon is also observed in speech, where non-native speakers tend to articulate more clearly than native speakers. Another major difference is the speed of signing - the experienced signers are considerably faster, which can be a challenge for the skeleton tracking algorithm.

For every sign either one hand is dominant and the other is secondary (this is not related to whether the signer is left- or right-handed) or the sign is fully symmetrical. The one hand dominance in non-symmetrical sings became obvious while studying the data from the experienced signers, where there was full consistency between signers in that respect. On the contrary, the inexperienced signers used

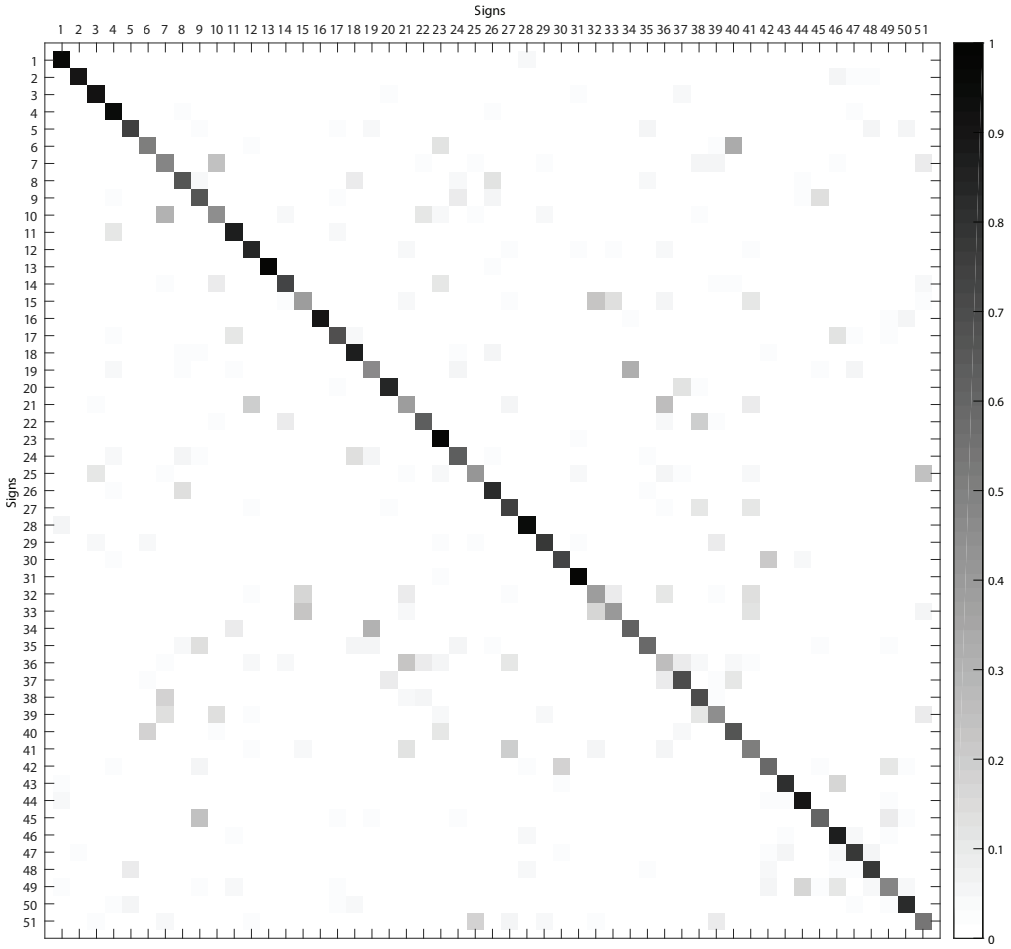


Figure 4: Confusion matrix for all inexperienced signers.

the hand that is dominant for writing, as dominant during signing. Since the target group of the system is not experienced in signing, it was decided to create models for both left- and right hand dominant versions of all signs, yielding an extended vocabulary of 102 signs, where the 51 signs in the original vocabulary are accompanied by their mirrored versions. During gameplay, once a sign is recognized to be performed *correctly* by the child but the dominant hand is swapped, we could provide feedback regarding this *mistake*.

The results obtained in these experiments show that the performance of the signer-independent recognizer is likely to be good enough for the target application. There are two main problems we plan to investigate. There is much room for improvement in the recognition accuracy. Examining the signer-dependent experiment, we can conclude that the simple spatial features used in this work are not sufficient. This is due to the fact that the skeleton tracker commits many errors in the estimates for the joints position, but also some signs are almost indistinguishable from spatial point of view (e.g. *yellow* and *orange*). We plan to extend the feature extraction procedure with a color-based hand tracker. The tracker is based on adaptive modeling of the skin color of the current user (by taking the face as reference). We expect that the hand tracker will increase the robustness of the tracking when combined with the skeleton tracker. Furthermore, we can introduce hand shape features. Further improvements are expected by introducing adaptation of the models based on a small set of signs from the target signer that could be collected during an enrollment/training phase in the game. We plan to continue recording new signs but creating a big set of isolated signs is a time consuming process.

Acknowledgments

The project Tivoli (Sign learning through games and playful interaction) is supported by the Swedish Post and Telecom Authority (PTS). We would like to thank Krister Schönström for the support during the experienced signers dataset recording.

References

ASL. 2006. http://www.bu.edu/asllrp/cslgr/.

G. Awad, J. Han, and A. Sutherland. 2006. A Unified System for Segmentation and Tracking of Face and Hands in Sign Language Recognition. In *International Conference on Pattern Recognition*, volume 1, pages 239–242.

BSL. 2010. http://www.bslcorpusproject.org/.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition*.

H. Cooper, B. Holt, and R. Bowden, 2011. *Sign Language Recognition*, pages 539–562. Springer London.

DGS. 2010. http://www.sign-lang.uni-hamburg.de/dgs-korpus/.

DICTA. 2012. http://www.dictasign.eu.

R. Feris, M. Turk, R. Raskar, K. Tan, and G. Ohashi. 2004. Exploiting Depth Discontinuities for Vision-based Fingerspelling Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–155.

K. Fujimura and X. Liu. 2006. Sign Recognition Using Depth Image Streams. In *International Conference on Automatic Face and Gesture Recognition*, FGR'06, pages 381–386. IEEE Computer Society.

K. Grobel and M. Assan. 1997. Isolated Sign Language Recognition Using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 162–167. IEEE.

R. Grzeszcuk, G. Bradski, M. H. Chu, and J. Y. Bouguet. 2000. Stereo Based Gesture Recognition Invariant to 3D Pose and Lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 826–833.

S. Hadfield and R. Bowden. 2012. Generalised Pose Estimation Using Depth. In *European Conference on Trends and Topics in Computer Vision*, ECCV'10, pages 312–325. Springer-Verlag.

S. Hong, N. A. Setiawan, and C. Lee, 2007. *Real-Time Vision Based Gesture Recognition for Human-Robot Interaction*, pages 493–500. Springer Berlin Heidelberg.

C.-L. Huang and W.-Y. Huang. 1998. Sign Language Recognition Using Model-based Tracking and a 3D Hopfield Neural Network. *Machine Vision and Applications*, 10(5):292–307.

K. Imagawa, S. Lu, and S. Igi. 1998. Color-based Hands Tracking System for Sign Language Recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 462–467.

T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. 2004. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *British Machine Vision Conference*.

M. W. Kadous. 1996. Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language. In *Workshop on the Integration of Gesture in Language and Speech*, pages 165–174.

J.-S. Kim, W. J., and Z. Bien. 1996. A Dynamic Gesture Recognition System for the Korean Sign Language (KSL). *IEEE Transactions on Systems, Man, and Cybernetics*, 26(2):354–359.

S. Mitra and T. Acharya. 2007. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(3):311–324.

R. Munoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato. 2008. Depth Silhouettes for Gesture Recognition. *Pattern Recognition Letters*, 29(3):319–329.

K. Murakami and H. Taguchi. 1991. Gesture Recognition Using Recurrent Neural Networks. In *Conference on Human Factors in Computing Systems*, CHI'91, pages 237–242. ACM.

L. Rabiner and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.

L. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.

S. S. Rautaray and A. Agrawal. 2015. Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey. *Artificial Intelligence Review*, 43(1):1–54.

J. Segen and S. Kumar. 1999. Shadow Gestures: 3D Hand Pose Estimation Using a Single Camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 485.

SIGNSPEAK. 2012. http://www.signspeak.eu.

SSL. 2009. http://www.ling.su.se/english/research/research-projects/sign-language.

T. Starner and A. Pentland. 1995. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. In *International Symposium on Computer Vision*, ISCV'95, pages 265–. IEEE Computer Society.

T. Starner, J. Weaver, and A. Pentland. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

C. Vogler and D. Metaxas. 1997. Adapting Hidden Markov Models for ASL Recognition by Using Three-dimensional Computer Vision Methods. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 1, pages 156–161.

C. Vogler and D. Metaxas. 1998. ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis. In *International Conference on Computer Vision*, pages 363–369.

C. Vogler and D. Metaxas. 1999. Parallel Hidden Markov Models for American Sign Language Recognition. In *International Conference on Computer Vision*, pages 116–122.

M. B. Waldron and S. Kim. 1995. Isolated ASL Sign Recognition System for Deaf Persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–271.

J. Yamato, J. Ohya, and K. Ishii. 1992. Recognizing Human Action in Time-sequential Images Using Hidden Markov Model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385.

M.-H. Yang, N. Ahuja, and M. Tabb. 2002. Extraction of 2D Motion Trajectories and Its Application to Hand Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1061–1074.

Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. 2011. American Sign Language Recognition with the Kinect. In *International Conference on Multimodal Interfaces*, ICMI'11, pages 279–286. ACM.

J. Zieren and K.-F. Kraiss. 2004. Non-intrusive Sign Language Recognition for Human-Computer Interaction. In *Symposium on Analysis, Design and Evaluation of Human Machine Systems*, page 27.

J. Zieren and K.-F. Kraiss, 2005. *Robust Person-Independent Visual Sign Language Recognition*, pages 520–528. Springer Berlin Heidelberg.

Barack Obama's pauses and gestures in humorous speeches

Costanza Navarretta
University of Copenhagen
Njalsgade 136, Copenhagen
costanza@hum.ku.dk

Abstract

The main aim of this paper is to investigate speech pauses and gestures as means to engage the audience and present the humorous message in an effective way. The data consist of two speeches by the USA president Barack Obama at the 2011 and 2016 Annual White House Correspondents' Association Dinner. The success of the message is measured in terms of the immediate audience response. The analysis of the multimodally annotated data indicates that silent speech pauses structure and emphasise the discourse, and often precede the audience response. Only few filled pauses occur in these speeches and they emphasise the speech segment which they follow or precede. We also found a highly significant correlation between Obama's speech pauses and audience response. Obama produces numerous head movements, facial expressions and hand gestures and their functions are related to both discourse content and structure. Characteristics for these speeches is that Obama points to individuals in the audience and often smiles and laughs. Audience response is equally frequent in the two events, and there are no significant changes in speech rate and frequency of head movements and facial expressions in the two speeches while Obama produced significantly more hand gestures in 2016 than in 2011. An analysis of the hand gestures produced by Barack Obama in two political speeches held at the United Nations in 2011 and 2016 confirms that the president produced significantly less communicative co-speech hand gestures during his speeches in 2011 than in 2016.

1 Introduction

This paper investigates Barack Obama's use of speech pauses and co-occurring gestures as means to engage the audience and present humorous message in an effective way. Gestures are in what follows non obstructive communicative body behaviours. Speech and gestures are closely related temporally and semantically in face-to-face communication (Kendon, 2004; McNeill, 2005), and they have multiple and sometimes co-occurring functions. In particular, gestures are important signals in interaction management (Allwood et al., 1992) and they contribute to the expression of the message's content (Kendon, 2004). Similarly, speech pauses are frequent in oral communication, and they have functions which are both related to the content and the structure of the discourse (Maclay and Osgood, 1959; Goldman-Eisler, 1968; Duncan and Fiske, 1977; Shriberg, 1994; Navarretta, 2016). The speech pauses, which we include in this study are silent pauses and filled pauses, the most common being *um*, *ah*, and *uh*. The gestures we address are head movements, facial expressions and hand gestures.

The data of the study consist of two speeches by the American president Barack Obama at the Annual White House Correspondents' Association Dinner. Obama has been recognised to be a capable and elegant speaker by the press and researchers. They have especially praised the lyrical content of his discourses and the ability with which he delivers them, inter alia (Cooper, 2011). The speeches at the Annual White House Correspondents' Association Dinners are different from other presidential speeches because the president, according to the tradition, mocks himself, his collaborators, political adversaries, and the press corps.

We address speech pauses and gestures as means to engage the audience and present humorous content in an effective way and measure the success of Obama's messages in terms of the immediate audience response in the form of cheers, laughter and/or applause. We have transcribed Obama's speech, annotated his gestures and marked audience's response, and performed qualitative and quantitative analyses on these annotations.

The paper is organised as follows. In section 2, we present background literature, then in section 3 we describe the data and the annotations. Section 4 contains an analysis of the annotated data followed by a discussion in sections 5. Finally, in section 6, we conclude and present future work.

2 Gestures and Speech Pauses

The communicative functions of co-speech gestures and speech pauses are several, and both gestures and pauses are multifunctional. Co-speech gestures contribute to the content and the structure of discourse (McNeill, 1992; Kendon, 2004), and they regulate the interaction as feedback and turn management signals (Allwood et al., 1992; Sacks et al., 1974; Allwood et al., 2007). Finally, they can show the attitudinal state of the speakers and their interlocutors.

Speech pauses are voluntary or involuntary signals, which help regulating the interaction (Duncan and Fiske, 1977; Clark and Fox-Tree, 2002) and can signal that the speakers are planning and structuring their message (Maclay and Osgood, 1959; Goldman-Eisler, 1968; Shriberg, 1994; Chafe, 1987). The presence of numerous speech pauses can also indicate that the speakers are talking about difficult concepts (Reynolds and Paivio, 1968; Rochester, 1973) or are looking for the appropriate word (lexical retrieval) (Krauss et al., 2000). Hirschberg and Nakatani (1998) investigate speech pauses in read and spontaneous English speech and conclude that pauses are often used as markers of discourse structure.

Studies of the relation between speech pauses and gestures point out that they are not only temporally but also functionally related (Boomer and Dittman, 1964; Butterworth and Hadar, 1989; McNeill, 1992; Kendon, 2004; Esposito et al., 2001). In particular, Esposito and Esposito (2011) find that speech pauses often co-occur with gesture holds in English and Italian spoken data and suggest that speech pauses and gesture holds have the same function of introducing new information. More specifically, speech pauses introduce new information in the verbal modality while gestural holds introduce new information in the non-verbal one. The presence of speech pauses and gestures has also been found to contribute to the perception of naturalness of software talking agents (Cassell et al., 1994; Cassell, 2000; Maatman et al., 2005; Rehm et al., 2008).

Some studies have focused specifically on the function of speech pauses in humorous contexts. Examples are the analysis of ungrammatical silent pauses and rate of articulation in the sitcom *Friends* by Quaglio (2009) and Bilá (2014). According to them, sitcoms exhibit features of both spoken and written discourse because they are based on written texts but are acted as spontaneous speech. Speech pauses and their timing in comedy have also been addressed by both researchers and comedians. Many of these studies propose that changes in speech rate and pauses preceding punch lines are common means in humorous discourse. However, corpus-based studies by Attardo and Pickering (2011) and Attardo et al. (2011) do not confirm these assumptions. In their data, pauses do not precede punch lines and the speech rate does not change in the humorous and non-humorous parts of the conversations. Attardo et al. (2011) find though that during humorous conversations, speakers smile and laugh more often than in non-humorous conversations.

Sankey (1998) and Oliver (2013) point out that pauses in comedy are not only means to structure and emphasise the discourse, but they also give the audience time to reflect on and appreciate the conveyed message. Finally, although the importance of gestures in comedy and film is recognised, gestures in humorous discourse are often analysed independently of speech (Clayton, 2007; Weitz, 2012).

The role of speech pauses and non-verbal behaviour as communicative means in both political speeches and in humorous discourse has also been addressed by various studies. For example, Duez (1982) analyses silent and non silent pauses in French casual interviews, political interviews and televised political speeches and she reports that the total time of silent pauses is 50% longer in political speeches than in interviews, and that the longer pauses often have a stylistic function. Salvati and Pettorini (2013) analyse different Italian speeches held by Silvio Berlusconi and conclude that he uses more emphatic pauses in political speeches than in other types of discourse. Guerini et al. (2013) collect a corpus of transcriptions of American political speeches and add to the transcriptions of the speeches occurrences of audience reaction in the form of Laughter and Applause in order to find prominent discourse segments in them. In the present work, we also annotate audience reaction, but differing from Guerini et al. (Guerini et al., 2013), we also annotate Obama's speech pauses and his gestures and investigate their relation to the audience's reaction.

3 The Speeches and the Annotations

The two speeches by the U.S. president Barack Obama at the White House Correspondents' Association Annual Dinner were held in 2011 and 2016, respectively. In the rest of the paper, we refer to the speech from 2011 as *talk2011* while we refer to the the speech from 2016 as *talk2016*. The videos which we have used are the official recordings by the White House which were available at `http:\\www.WH.gov` while Obama was the president. In the two videos, the president is recorded frontally while he speaks as it can be seen in the two snapshots in Figure 1 and 2.

We converted the recordings from mp4 to avi format and extracted audio wav files. Silent pauses were transcribed automatically from the audio applying a PRAAT built-in script (Boersma and Weenink, 2009). We found that the best silent threshold for delimiting silent pauses in these data is -35.0 dB with the minimum silent interval

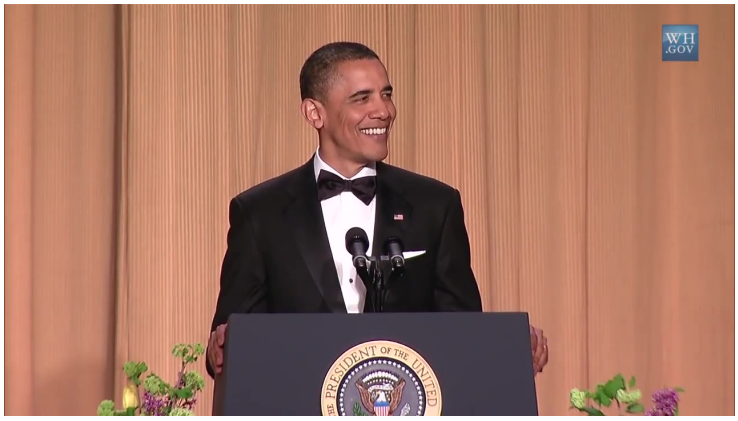


Figure 1: Snapshot from Obama's 2011 Speech

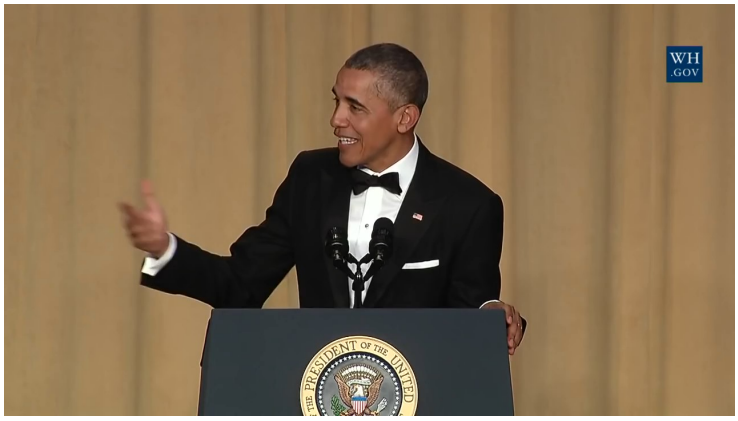


Figure 2: Snapshot from Obama's 2016 Speech

set at 0.2 seconds. Successively, Obama' s speech and audience reaction were added to the TextGrid file produced by PRAAT and the automatic transcriptions of silent pauses were corrected. The resulting speech transcriptions comprise speech segments (one or more speech tokens), silent and filled pauses, and, finally, audience responses (cheers, laughter and/or applause). The PRAAT transcriptions were imported in the ANVIL tool (Kipp, 2004) and Obama's gestures were annotated in it.

We only annotated the video parts in which Obama speaks and ignored the video segments which Obama shows to the audience. The duration of the annotated *talk2011* video segments is 13 minutes and 22 seconds while the duration of the annotated *talk2016* video segments is 30 minutes.

The shape and semiotic type of the gestures were annotated following the MUMIN annotation framework (Allwood et al., 2007). Table 1 shows the shape features of head movements, facial expressions and hand gestures.

Head movements are described by the form of the movement and information about whether the movement is single or repeated. The face is described by a general face attribute and three attributes describing the shape of eyebrows, the openness of the mouth and the position of the lips. Finally, hand gestures are characterised by the following information: a) the hand(s) involved in the gesture, b) whether the gesture is single or repeated, c) the trajectory of the hand(s), d) the extension of the fingers e) the orientation of the palm at the stroke. The semiotic types of the gestures which we annotated were inspired by Peirce (1931) and comprise *indexical deictic*, *indexical non-deictic*, *iconic*, *iconic metaphoric* and *symbolic* (Allwood et al., 2007).

4 The Analysis

Table 2 contains the absolute and relative frequency (occurrence per second) of speech tokens, pauses and gestures in *talk2011*, *talk2016*, and in total. When calculating the ratio speech token per second, we excluded the time during which Obama does not speak because the audience is laughing and/or applauding. The resulting speech duration is 8 minutes for *talk2011* and 19 minutes for *talk2016*.

Table 1: Shape Features

Attribute	**Value**
HeadMovement	Nod, Jerk, HeadForward, HeadBackward, Shake, Waggle, HeadOther, Tilt, SideTurn
HeadRepetition	HeadSimple, HeadRepeated
General face	Smile, Laugh, Scowl, FaceOther
Eyebrows	Raise, EyebrowsOther
MouthOpen	OpenMouth, CloseMouth
MouthLips	CornersUp, CornersDown, Protruded, Retracted, LipsOther
Handedness	BothHandsSym, BothHandsAsym, RightSingleHand, LeftSingleHand
HandRepetition	Single, Repeated
Fingers	IndexExtended, ThumbExtended, AllFingersExtended, FingersOther
TrajectoryLeftHand	LeftHandForward, LeftHandBackward, LeftHandSide, LeftHandUp, LeftHandDown, LeftHandComplex, LeftHandOther
TrajectoryRightHand	RightHandForward, RightHandBackward, RightHandSide, RightHandUp, RightHandDown, RightHandComplex, RightHandOther
PalmOrientation	PalmUp, PalmDown, PalmSide, PalmVertical, PalmOther

Table 2: Absolute and Relative Frequency of Speech, Pauses and Gestures

token	**talk2011 #**	**talk2011 #/sec**	**talk2016 #**	**talk2016 #/sec**	**Total**	**Total/sec**
speech	1059	2.21	2531	2.22	3590	2.22
silent	225	0.47	243	0.21	468	0.29
filled	10	0.02	10	0.009	20	0.01
head	357	0.74	831	0.72	1188	0.73
face	66	0.14	117	0.16	183	0.15
hand	51	0.11	237	0.21	289	0.18

In order to verify whether Obama behaves differently in the two events in terms of speech rate and frequency of gestures, we tested the χ square of the observed and expected behaviour, the expected behaviour being that Obama speaks or produces a type of gesture with the same relative frequency in the two events. The difference between observed and expected behaviour is considered to be significant if the χ square p value is less than 0.001.

There is no difference in speech rate (words and fillers) in the two speeches, and this could be related to the fact that the two speeches are read an they belong to a particular genre. In *talk2011*, silent and filled pauses occur more frequently than in *talk2016*, but the difference is not statistically significant (χ square $= 2.77$ with 1 degrees of freedom and the 2-tailed $p = 0.096$). Silent pauses have shorter duration in *talk2011* than in *talk2016*, but also in this case the difference is not significant.

The low frequency of fillers in the two speeches was expected because Obama is reading from a manuscript. The qualitative analysis of the fillers' occurrences shows that Obama uses them consciously to give importance to what he has said or what he is going to say. In a case in *talk2016*, after having made a joke on the journalists' exaggerate coverage of Trump's campaign, Obama repeats three times the filler *hm*, and then after a longer silent pauses utters a strong *hm* in order to emphasise his point giving rise to the audience laughter.

Obama moves his head with approximately the same frequency in *talk2011* and *talk2016* (χ square equals 0.101 with 1 degrees of freedom and the 2-tailed $p = 0.751$). He moves his head continuously turning it to the right or to the left in order to address the whole audience. Sometimes, he also turns his body and talks or refers to the organisers and/or his wife who sit at the right of his podium. During speech pauses, or while the audience is laughing, Obama often moves his head forward to look at his manuscript. The latter head movements are not communicative. It must be noted that the co-occurring upper body movements were not annotated.

Also the relative frequency of facial expressions in *talk2011* and *talk2016* is approximately the same ($\chi square = 0.71$ with 1 degrees of freedom, and the 2-tailed $p = 0.4$).

The most common facial expressions in these data are smiles, laughs and expressions in which Obama retracts his lips while listening at the audience's response or looking at his manuscript. Frowns and raised eyebrows also occur frequently in these data.

While the frequency of Obama's head movements and facial expressions does not change in the two talks, we found that he produces significantly more hand gestures in *talk2016* than in *talk2011* (χ square equals 19.295 with

1 df, and 2-tailed $p < 0.0001$). All types of co-speech hand gesture are produced in the two speeches, and the must frequent gesture types are beats and deictics. Hand gestures are also used to structure and emphasise the discourse. There are relatively more symbolic gestures in *talk2011* than in *talk2016*, and Obama produces relatively more iconic and deictic gestures in *talk2016* than in *talk2011*.

The temporal ratio of audience reaction and Obama's speech is approximately the same in the two events. More specifically, the audience applauds and/or laughs 37% of the speech duration in *talk2011* and 40% of the speech duration in *talk2016*. The relative frequency of response per second in *talk2011* is 0.18 while in *talk2016* is 0.15. Thus, the audience laughed/applauded more often, but for shorter time in *talk2011* than they did in *talk2016*, but the difference in response frequency is not statistically significant (χ square equals 2.686 with 1 degrees of freedom, and 2-tailed $p = 0.1012$).

In some cases, Obama repeats words in order to emphasise them or prolong the audience reaction. An example of this is in *talk2011* in which he makes fun of his wife's engagement in health food and children. It is illustrated in Table 3 in which three columns we report speech segments and pauses, co-occurring gestures and audience reaction.

Table 3: An Example from Talk2011

Obama's speech	**Co-speech gestures**	**Audience**
We made a terrific team	Turns right and points at Michelle	
at the Easter Egg Roll this week [Pause]	Turns to the front	
I'd give out [Pause]		
bags of candy to the kids and [Pause]	Performs handing gesture with right hand (Figure 3a)	
and she'd snatch them right back		
out of their little hands [Pause]	Both hands illustrate the snatching (Figure 3b)	Audience laughter
laughs [Pause]	Turns to the right and looks at Michelle	Audience laughter
laughs	Turns to the front	Audience laughter
Snatched them! [Pause]	Smiles	Audience laughter

Figure 3: Snapshots of Iconic Gestures: a) Handing a Bag b) Snatching Bags

In Table 4, an example of uses of silent pauses as emphasising signals and means to get the audience react to the speech content is given. The example is from *talk2016* at a point where Obama makes fun of Hillary Clinton's clumsiness in using the social media. Table 4 is organised as the preceding Table 3.

Finally, we measured the Pearson correlation coefficient between pause occurrences and audience response in order to confirm that pauses are related to audience reaction in these data. The correlation is positive and is highest when only silent pauses were considered (Pearson 2-tailed correlation $r = 0.465$). The correlation level is also highly significant ($r(1541) < 0.0001$).

Table 4: Example from Talk2016

Obama's speech	**Co-speech gestures**	**Audience**
You've got to admit though [Pause]		
Hillary trying to appeal to young voters [Pause]		
is a little bit like your relative [Pause]		
who just signed for Facebook [Pause]		Audience laughter
Dear America [Pause]		
did you get my poke? [Pause]	Smiles	Audience laughter
Is it appearing on your wall? [Pause]	Repeatedly moves both hands one up and one down	
	(Figure 4)	Audience laughter

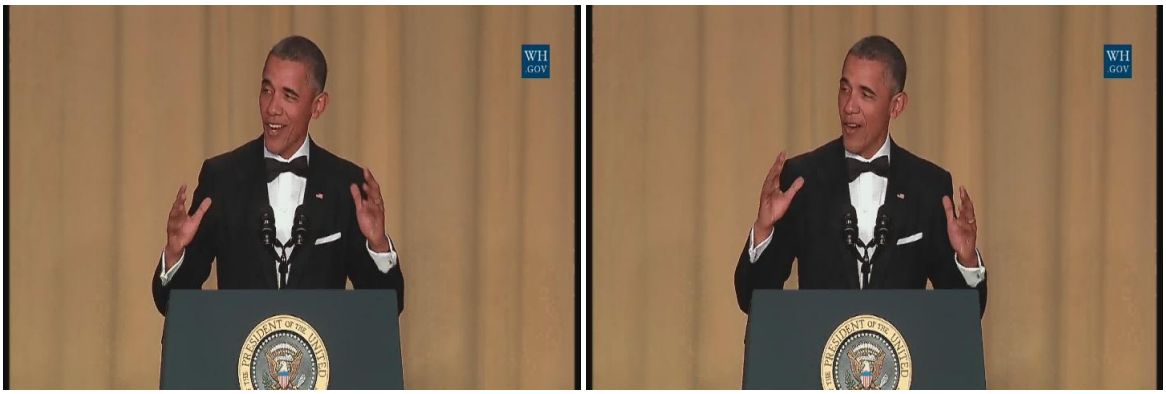


Figure 4: Snapshots of Iconic Gesture Co-occurring with Speech Segment "appearing on your wall" from *talk2016*

5 Discussion

The analysis of the multimodally annotated Obama's speeches at the Annual White House Correspondents' Association Dinner in 2011 and 2016 shows that Obama produces slightly more silent pauses in *talk2011* than in *talk2016*. The audience applauds more often in *talk2011* than in *talk2016* while they applaud/laugh for a longer time during the latter event. However, these differences are not statistically significant.

The analysis of silent pauses in the two speeches indicates that pauses, in most cases, delimit grammatical phrases, that is nominal, adjectival, verbal, adverbial and clausal phrases. Obama also delimits topic shifts with silent pauses, thus these data confirm what Hirschberg and Nakatani (1998) found in various types of read and spoken English data. We also found a number of pauses which preceded single words. Since Obama is mostly reading from a manuscript, these pauses cannot signal lexical retrieval as it would be the case in spontaneous speech (Krauss et al., 2000). Furthermore, they are not used to fake spontaneous speech as it is the case in sitcoms (Quaglio, 2009; Bilá, 2014). Instead, they are so-called *emphatic pauses* and emphasise the following speech segment. Finally, a number of pauses that follow a word or a phrase are used by Obama to let the audience get the point, and in numerous cases after these pauses the audience react by laughing and/or applauding the president. These emphasising use of pauses in humorous speech are described inter alia by Sankey (1998) and Oliver (2013). The correlation between pauses and audience response is positive as we expected, and therefore information about pauses should be tested as a feature for predicting audience reaction to humorous speech.

Filled pauses are few in these data, and Obama uses them consciously to emphasise what he has just said as in the case in which he repeats the same filler several times making the audience laugh. Finally, a number of short pauses follow the audience response, indicating that Obama adapts his speech to the audience's reaction. In some cases, he controls the duration of the audience's response by e.g. signalling with his hands that they should stop laughing/applauding or in other ways indicating that he wants to talk. In other cases, Obama makes the audience continue laughing by laughing with them or repeating parts of his preceding words as in the example illustrated in Table 3.

We did not compare Obama's speech rate in humorous and not humorous speech and therefore cannot confirm Attardo and Pickering (2011)'s observations, but our data confirms the findings by Attardo et al. (2011) that noticed that people often laugh and smile when they deliver jokes. This was also the case for Obama who often laughs and

smiles in these speeches.

Obama often moves his head to look at his manuscript while he holds speech pauses. These gestures are not communicative, but they are typical of people who are used to present read material in a lively way. Communicative gestures only seldom co-occur with speech pauses, and this confirms the findings by Esposito and Esposito (2011) who report that gestural holds often co-occur with silent pauses and have parallel functions of introducing new gestural and verbal content, respectively.

Obama interacts continuously with his audience, talking directly and pointing to individuals in the room. Sometimes, he laughs while talking and this is often the case when he presents jokes involving his wife, Michelle, his collaborators or political adversaries. Therefore, even if many of the head movements and hand gestures produced by Obama are typical of read speeches, the many occurrences of laughter, smiles and deictics (hand gestures and head movements) pointing to people in the room, are behaviours which are not typical of political speeches. Interestingly, Obama varies the way in which he presents his jokes. Sometimes, he laughs while talking, other times he is extremely serious and therefore surprises his audience who starts laughing when they find out that he was ironic or said a joke. In these cases, Obama stops talking in order to allow the audience to understand the point. This effect of surprise in the presentation of humorous speech has been discussed inter alia by Beeman (1999).

Concluding, Obama uses both speech pauses and gestures to present his humorous speech and combines behaviours from normal political speeches with behaviours typical of humorous discourse. We did not find differences in speech rate and frequency of head movements and facial expressions in the two speeches, but we found significant differences in the relative frequency of hand gestures. In fact Obama moves his hands significantly more often in *talk2016* than in *talk2011*. In order to determine whether this difference in hand gesturing is a special case or reflects a development in the way Obama's presents his speeches, we annotated the communicative hand gestures in the first five minutes of two political speeches, which Obama held at the United Nations in 2011 and in 2016.[1] In the 2011 speech Obama produced 26 hand gestures (0.08 hand gestures per second) while in the 2016 he produced 68 hand gestures (0.23 hand gestures per second). The difference in occurrence frequency is statistically significant (χ square equals 18.766 with 1 df, and two tailed $p < 0.0001$) and confirms the difference in the frequency of hand gestures which we have discovered in *talk2011 talk2016*. Concluding, the two sets of Obama's speeches, which we have analysed, show clearly that Obama uses more frequently co-speech gestures in 2016 than in 2011. This indicates that Obama improves his presentation technique during his presidential term.

6 Conclusions

In this paper, we have presented a study of pauses and communicative gestures in annotated audio- and video-recordings of two humorous speeches by the American president Obama at the Annual White House Correspondents' Association Dinners in 2011 and 2016 with the aim of determining to what extent pauses and gestures contribute to the successful presentation of humorous discourse. Success of presentation was measured in terms of direct audience response. Silent pauses were automatically extracted from the audio files in PRAAT, Obama's speech and the audience reaction were transcribed and Obama's head movements, facial expressions and hand gestures were annotated.

The analysis of the data shows that Obama's speech rate and the relative frequency of head movements and facial expressions is the same in 2011 and 2016, while Obama produces significantly more hand gestures in 2016 than in 2011. An analysis of the hand gestures produced by Obama in the first part of two political speeches held at the United Nations in the same years confirms that Obama uses more frequently co-speech hand gestures in the last year of his presidency than five years earlier. Since hand gestures contribute to the presentation of the content and structure of discourse, it is evident that Obama improves his presentation technique during his two presidential terms.

Obama uses all kinds of pauses and gestures, and especially facial expressions, as means for emphasising and presenting effectively his discourse.

We found that audience responses are related to silent pauses in these data and that the correlation is significant. Therefore, we are currently testing to what extent information about silent pauses contributes to the prediction of audience response in the speeches.

Finally, it must be noted that we have not included speech content and intonation features in our analysis. These are central aspects of humorous speech, and should therefore be accounted for in the future. The type of audience, and the communicative situation are also relevant with respect to the audience reaction to jokes. These aspects should also be included in future investigations of humorous discourse.

[1]The two speeches are available at`https://www.youtube.com/watch?v=UK7JEYqIfw4` and `https://www.youtube.com/watch?v=ji6pl5Vwrvk`, respectively. We analysed 306 seconds of the 2011 talk and 300 seconds of the 2016 talk excluding sequences of video frames in which the audience reaction is recorded and Obama's hand gestures are therefore not visible.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing. *Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation*, 41(3–4):273–287.

Salvatore Attardo and Lucy Pickering. 2011. Timing in the performance of jokes. *Humor - International Journal of Humor Research*, 24(2):233–250.

Salvatore Attardo, Lucy Pickering, and Amanda Baker. 2011. Prosodic and Multimodal Markers of Humor in Conversation. *Pragmatics and Cognition*, 19(2):224–247.

William O. Beeman. 1999. Humor. *Journal of Linguistic Anthropology*, 9(1/2):103–106.

Magdalena Bilá. 2014. A Comparative Analysis of Silent Pauses and Rate of Articulation in the Discourse of Sitcom. *Discourse and Interaction*, 7(1).

Paul Boersma and David Weenink, 2009. *Praat: doing phonetics by computer*. Retrieved May 1, 2009, from http://www.praat.org/.

D.S. Boomer and A.T. Dittman. 1964. Speech rate, filled pause and body movement in interviews. *Journal of Nervous and Mental Disease*, 139:324–327.

B.L. Butterworth and U. Hadar. 1989. Gesture, speech, and computational stages: A reply to mcneill. In *Psychological Review*, volume 96, pages 168–174.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Justine Cassell. 2000. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78.

W.L. Chafe. 1987. Cognitive Constraint on Information Flow. In R R. Tomlin, editor, *Coherence and Grounding in Discourse*, pages 20–51. John Benjamins, Amsterdam.

Hebert H. Clark and Jean E. Fox-Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84:73–11.

Alex Clayton. 2007. *The body in Hollywood slapstick*. McFarland.

Marilyn M. Cooper. 2011. Rhetorical agency as emergent and enacted. *College Composition and Communication*, 62(3):420–449.

Danielle Duez. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech*, 25(1):11–28.

S. Duncan and D.W. Fiske. 1977. *Face-to-face interaction*. Erlbaum, Hillsdale, NJ.

Anna Esposito and Antonietta M. Esposito. 2011. On Speech and Gesture Synchrony. In Anna Esposito, Alessandro Vinciarelli, Klara Vicsi, Catherine Pelachaud, and Anton Nijholt, editors, *Communication and Enactment - The Processing Issues*, volume 6800 of *LNCS*, pages 252–272. Springer-Verlag.

Anna Esposito, Karl Erik McCullough, and Frank Quek. 2001. Disfluencies in gesture: gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE International Workshop on Cues in Communication*, Hawai.

Frieda Goldman-Eisler. 1968. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, London.

Marco Guerini, Danilo Giampiccolo, Giovanni, Moretti, , Rachele Sprugnoli, and Carlo Strapparava. 2013. The new release of corps: A corpus of political speeches annotated with audience reactions. In Isabella Poggi, Francesca D'Errico, Laura Vincze, and Alessandro Vinciarelli, editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action: International Workshop, Political Speech 2010, Rome, Italy, November 10-12, 2010, Revised Selected Papers*, pages 86–98. Springer Berlin Heidelberg, Berlin, Heidelberg.

J. Hirschberg and C. Nakatani. 1998. Acoustic Indicators of Topic Segmentation. In *Proceedings of ICSLP-98*, Sidney.

Adam Kendon. 2004. *Gesture - Visible Action as Utterance* . Cambridge University Press.

Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Ph.D. thesis, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com.

R.M Krauss, Y. Chen, and R. F. Gottesman. 2000. Lexical gestures and lexical access: a process model. In D. McNeill, editor, *Language and gesture*, pages 261–283. Cambridge University Press.

RM Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural behavior of a listening agent. In *Intelligent Virtual Agents*, pages 25–36. Springer.

Howard Maclay and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous english speech. *Word*, 15:19–44.

D. McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press, Chicago.

David McNeill. 2005. *Gesture and thought*. University of Chicago Press.

Costanza Navarretta. 2016. The functions of fillers, filled pauses and co-occurring gestures in danish dyadic conversations. In Linköping University Electronic Press, editor, *Postproceedings of the 3rd European Symposium on Multimodal Communication*, volume 105, pages 55–61.

Bobbie Oliver. 2013. *The Tao of Comedy: Embrace the Pause*. Oliver.

C. S. Peirce. 1931. *Collected Papers of Charles Sanders Peirce, 1931-1958, 8 vols*. Harvard University Press, Cambridge, MA.

P. Quaglio. 2009. *Television Dialogue. The sitcom Friends vs. natural conversation*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Matthias Rehm, Yukiko Nakano, Elisabeth André, and Toyoaki Nishida. 2008. Culture-specific first meeting encounters between virtual agents. In *Intelligent virtual agents*, pages 223–236. Springer.

Allan Reynolds and Allan Paivio. 1968. Cognitive and emotional determinants of speech. *Canadian Journal of Psychology*, 22:164–175.

Sherry R. Rochester. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2:51–81.

H. Sacks, E. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Luisa Salvati and Massimo Pettorino. 2013. A Diachronic Analysis of Face-to-Face Discussions: Berlusconi, Fifteen Years Later. In Isabella Poggi, Francesca D'Errico, Laura Vincze, and Alessandro Vinciarelli, editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action: International Workshop, Political Speech 2010, Rome, Italy, November 10-12, 2010, Revised Selected Papers*, pages 65–74. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jay Sankey. 1998. *Zen and the Art of Stand-Up Comedy*. Routledge, New York.

Elisabeth Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of California, Berkeley.

Eric Weitz. 2012. Failure as success: On clowns and laughing bodies. *Performance Research*, 17(1):79–87.

Identification of Emphasised Regions in Audio-Visual Presentations

Keith Curtis
ADAPT Centre
School of Computing
Dublin City University
Ireland
Keith.Curtis
@AdaptCentre.ie

Gareth J.F. Jones
ADAPT Centre
School of Computing
Dublin City University
Ireland
Gareth.Jones
@dcu.ie

Nick Campbell
ADAPT Centre
School of Computer
Science & Statistics
Trinity College Dublin
Ireland
nick@tcd.ie

Abstract

Rapidly expanding archives of audio-visual recordings available online are making unprecedented amounts of information available in many applications. New and efficient techniques to access this information are needed to fully realise the potential of these archives. We investigate the identification of areas of intentional or unintentional emphasis during audio-visual presentations and lectures. We find that, unlike in audio-only recordings where emphasis can be located using pitch information alone, perceived emphasis can be very much associated with information from the visual stream such as gesticulation. We also investigate potential correlations between emphasised speech, and increased levels of audience engagement during audio-visual presentations and lectures.

1 Introduction

The rapidly expanding archives of audio-visual recordings available online are making unprecedented amounts of information available in many applications. However, realising the potential of this content requires the development of new and innovative tools to enable the efficient location of significant content of interest to the user. Manually browsing multimedia archives to identify audio-visual content of interest is extremely time consuming. While browsing of this sort of content is challenging for multimedia archives in general, it is an extremely challenging problem for archives of spoken content where most of the information exists in the audio stream.

Addressing requires the development of new tools to allow the user to search for potential content of interest without having to first listen to it. In our current work we are interested in identification of areas of speaker emphasis in audio-visual presentations. This has the potential to improve applications such as automatic summarisation or browsing of audio-visual content. Tools of this nature could also potentially be used for improved search and retrieval capabilities.

Previous work has explored identification of emphasised speech using the audio-only stream. In this work we expand on this earlier work in an audio-visual context to demonstrate that emphasis detection can be more successfully achieved using a multimodal analysis approach.

We also address the question of whether speech emphasis in the context of audio-visual presentations or lectures shows a correlation with what is typically referred to as 'good' public speaking techniques. This concept of 'good' public speaking techniques has been investigated in our previous work (Curtis et al., 2015). Given that emphasis is normally applied by the speaker to draw the attention of the audience to a specific part of speech for reasons of clarity or importance, we also investigate whether this applied emphasis affects change in the overall levels of engagement among the audience to such material.

2 Previous Work

Previous work (Chen and Withgott, 1992) has studied the use of emphasis for automatic summarisation of a spoken discourse. In this work emphasised speech from one speaker was detected and summarisation excerpts were extracted with no noticeable differences from human extracted summarisation excerpts. The data used was a 27 minutes videotaped interview between two primary speakers and the second was a set of phrases extracted from a telephone conversation. The emphasis model was trained on a

Hidden Markov Model (HMM) in which three separate models were created for 3 speech emphasis levels: emphatic speech, unemphatic speech and background speakers.

Another study (Arons, 1994) performed pitch based emphasis detection for automatic segmentation of speech recordings. In this work a pitch threshold of the top 1% of pitch values was chosen, speech segments with pitch values exceeding this threshold were classified as emphasised speech. From this, the pitch based segmentation technique was used to summarise the speech recordings into the most important speech segments. (He et al., 1999) attempted to summarise audio-visual presentations using pitch values in the top 1 percentile. In this work they found that audio-visual presentations were less susceptible to pitch based emphasis analysis than the audio stream only.

Following on from this work, (Kennedy and Ellis, 2003) studied emphasis detection for characterisation of meeting recordings. In this work 5 human annotators labelled 22 minutes of audio from the International Computer Science Institute (ICSI) meeting corpus. Annotators were given both an audio recording and a transcript from the meeting. Annotators listened to the audio recording while working their way down the transcript and marking each utterance as emphasised or not. They extracted pitch and aperiodicity of each frame and calculated the mean and standard deviation for each speaker. In cases where 4 or more human annotators agreed on emphasis accuracy rates of 92% were achieved. In addition, the utterances found to be the most emphasised were found by annotators to be a good summarisation of the meeting recording.

To the best of our knowledge, the detection of regions of speech emphasis has not previously been performed in an audio-visual context. (He et al., 1999) indicate that emphasis in audio-visual recordings is indicated by more than just notable increases in pitch as in the audio stream. In this study we investigate use of audio-visual features to detect emphasis in academic presentations. Also to the best of our knowledge, this concept has not been investigated for potential correlations with resulting audience engagement to the presentation or lecture material at hand.

3 Multimodal DataSet

For this study we used the International Speech Conference Multi-modal Corpus (SCMC) developed in our previous work (Curtis et al., 2015). This contains 31 academic presentations totalling 520 minutes of video, and includes high quality 1080p parallel video recordings for both the speaker and the audience to each presentation. Recordings have a frame rate of 29.97 fps, and were recorded at H264 codec. High quality parallel audio recordings are also included for each presentation in addition to close-up recordings of each presenter's slides. The majority of presentations are standard podium presentations by a single speaker in front of a seated audience. Two of the presentations consisted of podium presentations by two speakers in front of their seated audience. For this study, four presentations were selected from the corpus, two of which had male presenters and two with female presenters, each presentation was in English.

To limit the size of this preliminary investigation, a single 5-minute clip was selected from each presentation, totalling 20 minutes of presentation video used in this initial study. Segments were chosen to include presenters who were judged by human annotators in previous work on this dataset to be good presenters (Curtis et al., 2015), and to exclude regions of speech not of the presenter.

4 Multimodal Feature Extraction

For our investigation we extracted the following audio-visual features from the recorded presentations:

Pitch: AutoBi Pitch Extractor (Rosenberg, 2010). We use default min and max values of 50 and 400 respectively. Pitch values over the entire range were normalised for each speaker.

Intensity: AutoBi Intensity Extractor (Rosenberg, 2010). This generated an Intensity contour using default parameters of a minimum intensity of 75dB and a timestep of 100ms. Intensity values over the entire range were normalised for each speaker.

Head movement: OpenCV (Bradski and Kaehler, 2008) using Robust Facial Detection described in (Viola and Jones, 2004). For this task we used a Head and Shoulder cascade to detect the presenters head and return the pixel values for the location of the speakers head at that point in time. We then extracted

Figure 1: Presenter: mid-Emphasis

head movement by taking the Euclidean distance between pixel points in corresponding frames. These values were then normalised for each speaker.

Speaker Motion: was extracted using an optical flow implementation in OpenCV described in (Lucas et al., 1981). We calculated the total pixel motion changes from frame to frame to put more weight on directional changes in motion and take the mean and standard deviation in overall speaker motion. This accounted for change of direction in motion and represented variances of these values. These values were normalised per speaker.

5 Experimental Investigation

This section explores the experimental investigation we undertook during this research, including the human annotation of speaker emphasis during academic presentations.

5.1 Initial Investigation

The first part of this investigation involved asking a total of 10 human annotators to watch 2 of the 5 minute video clips taken from the four presentations, totalling 10 minutes of presentation video to be annotated for this concept. The annotators were asked to mark areas where they considered the presenter to be giving emphasis to speech, either intentionally or unintentionally. Due to the subjectivity of this task, annotators were instructed beforehand as to what exactly constituted emphasised speech, and were allowed to decide themselves just what they considered to be emphasised. There was however much disagreement between human annotators over areas of emphasis. We consider this to be due to the high level of subjectivity on just what it means to emphasise. This high-level of disagreement meant it was not practical to train a machine learning algorithm over this data for automatic classification of emphasised speech.

5.2 Further Investigation

Because of this disagreement, and in order to better understand the characteristics of regions consistently labelled as emphasised speech, we studied areas of agreed emphasis between the annotators. It was clear from this analysis, that consistent with earlier work, all agreed upon areas of emphasis occur during areas of high pitch, but also in regions of high visual motion coinciding with an increase in pitch. Following this an extraction algorithm was developed using the features listed in the previous section to locate further candidate areas of emphasis.

The algorithm selects candidate regions by finding areas of high pitch in combination with areas of high motion or head movement. A two second gap was allowed between areas of high pitch and high movement on the part of the speaker for selection of areas of emphasis. Candidate emphasised regions were marked from extracted areas of pitch within the top 1, 5, and top 20 percentile of pitch values, in addition to the top 20 percentile of gesticulation down to the top 40 percentile of values respectively. This resulted in the extraction of 83 candidate areas of emphasised speech from our dataset. These candidate

regions were each judged for emphasis by three separate human judges, with the majority vote on each candidate emphasis region taken as the gold standard label for final agreement of emphasis.

6 Analysis and Results

Of the 83 candidate areas of emphasis extracted from presentation segments, 18 had pitch values in the top 1 percentile after normalisation. Of these 18 candidate areas, four were accompanied by speaker motion, mostly gesturing, sometimes head movement, while 14 were not accompanied by any speaker movement or gesturing of any significance. All of the 4 candidate areas accompanied by movement or gesturing were judged by human annotators to be emphasised regions of speech. Only 5 of the 14 candidate areas not accompanied by gesturing or movement of any sort were judged by human annotators to be emphasised speech. This indicates that in audio-visual context, emphasised speech frequently depends on gesturing and / or other movement in addition to pitch.

Fifteen of the candidate areas of emphasis were in the top 5 percentile of pitch values extracted. Three of these were accompanied by gesturing on the part of the presenter. All three of these areas accompanied by gesturing were judged by human annotators to be emphasised speech. Of the 12 areas not accompanied by any gesturing by the presenter, only 5 were judged to be emphasised by our human annotators. A total of 33 emphasis candidates were extracted from pitch values in the top 5 percentile. Seven were accompanied by gesturing and all of these were judged by human annotators to be emphasised. Twenty-six were not accompanied by gesturing, and only 10 of these were judged by the human annotators to be emphasised. It was found that candidate emphasis regions in the top 20 percentile of pitch values and the top 20 percentile of gesticulation combined were true regions of emphasis as labelled by our human annotators. The mean intra-class correlation was calculated as 0.5818, giving us a good level of inter annotator agreement between judges.

As the examples used thus far provided very few samples to definitively state reliable results, we extracted 15 additional samples of emphasised speech from the corpus. These were extracted from areas where normalised motion and pitch both exceed the top 20 percentile with a two-second gap. In addition, Thirteen additional samples of non-emphasised speech were used. Three additional human annotators were recruited to annotate new candidate emphasis area. Thirteen of the 15 emphasised areas were labelled by human annotators as emphasised speech.

As indicated by the above results, all annotated areas of emphasis contain significant gesturing in addition to pitch with the top 20 percentile. Gesturing was also found to take place in non-emphasised parts of speech, however this was much more casual and not accompanied by pitch in the top 20 percentile.

7 Correlations Between Speaker Rankings and Emphasised Speech

We calculate prospective correlations between annotated speaker ratings and annotated emphasised speech. To achieve this we take values for 4 separate 5 minute video clips containing original emphasis annotations. We achieve this by first calculating the average speaker rating for each 90 second time window, then summing the total number of emphasis detections within that time-frame. Time-windows are incremented at each step by 30 seconds.

Calculating this over all of the 5 minute video clips combined gives a total of 32 time-windowed instances. We calculate correlations using the Pearsons Correlation Coefficient Calculator. Following this, we also calculate the correlation for speaker specific correlations between speaker ratings and emphasised speech. Table 1 outlines the results of these tests.

Table 1: Speaker Ratings - Emphasis : Linear Correlation

Video	r =
All_Combined	-0.3247
plenaryoral_2	-0.2988
plenaryoral_11	-0.0845
plenaryoral_12	-0.3362
prp_2	0.7976

Although the calculation for all videos combined shows a weak but nonetheless existent negative correlation between speaker ratings and emphasis, when we look at the calculations for all videos we see that video prp_2 alone holds a strong positive correlation of 0.7976. With all other videos in the set showing a weak negative correlation, we can conclude that no true correlation exists between speaker ratings and emphasis.

8 Correlations Between Audience Engagement Levels and Emphasised Speech

We also calculate prospective correlations between annotated audience engagement levels and annotated emphasised speech. Once again, to achieve this we take emphasis values for 4 separate 5 minute video clips containing original emphasis annotations. We first calculate the average engagement level for each 90 second time window, then summing the total number of emphasis detections within that time-frame. Time-windows are incremented at each step by 30 seconds.

Calculating this over all of the 5 minute video clips combined gives a total of 32 time-windowed instances. Correlations are calculated using the Pearsons Correlation Coefficient Calculator. Following this, we also calculate the correlation for speaker specific correlations between audience engagement levels and emphasised speech. Table 2 shows the results, of which no clear correlation between these two concepts is visible.

Table 2: Audience Engagement - Emphasis : Linear Correlation

Video	r =
All_Combined	-0.1593
plenaryoral_2	-0.475
plenaryoral_11	0.2887
plenaryoral_12	0.8868
prp_2	0.1857

From Table 2 we can clearly see that correlation calculations per video appear to be very random, leading us to conclude that no correlations exist between audience engagement levels and emphasised speech. While the video plenaryoral_12 indicates a strong positive correlation, plenaryoral_2 indicates a medium negative correlation while other videos show no real correlation. Overall with no clear pattern emerging we can conclude that no correlation exists. However, it should of course be noted that this analysis is carried out over a very small set of data.

9 Conclusions and Further Work

Previous work on emphasis detection in recordings of spoken content had looked at the concept in the context of the audio-stream only. Our small study shows that emphasis of speech in the audio-visual stream very much depends upon speaker gesticulation in addition to pitch. However, speech intensity levels did not show any significant correlation with emphasis. These results demonstrate the importance of gesturing for emphasis in the audio-visual stream. Further, no real correlations were discovered between areas of 'good' public speaking techniques or with audience engagement levels.

Previous work had discovered that emphasised speech can be used for effective summarisation of the audio-only stream (Chen and Withgott, 1992). Our future work will investigate the potential to sum-

marise audio-visual lectures and presentations by using identified areas of intentional or unintentional speaker emphasis in addition to other paralinguistic features.

In this regard, initial experiments investigating the potential of identified areas of emphasised speech to be used for generating automatic presentation summaries have proven promising. In work combining identified areas of emphasised speech along with classifications for audience engagement and comprehension, early results have shown that generated summaries tend to be more engaging and information rich than full presentations, whilst participants tend to maintain focus for longer periods (Curtis et al., 2017).

10 Acknowledgments

This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Dublin City University. We would like to thank all human annotators for participating in this study.

References

Barry Arons. 1994. Pitch-based emphasis detection for segmenting speech recordings. In *International Conference on Spoken Langauge Processing*.

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.

Francine R Chen and Margaret Withgott. 1992. The use of emphasis to automatically summarize a spoken discourse. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 229–232. IEEE.

Keith Curtis, Gareth JF Jones, and Nick Campbell. 2015. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 35–42. ACM.

Keith Curtis, Gareth JF Jones, and Nick Campbell. 2017. Utilising high-level features in summarisation of academic presentations. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 315–321. ACM.

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pages 489–498. ACM.

Lyndon S Kennedy and Daniel PW Ellis. 2003. Pitch-based emphasis detection for characterization of meeting recordings. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 243–248. IEEE.

Bruce D Lucas, Takeo Kanade, et al. 1981. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, volume 81, pages 674–679.

Andrew Rosenberg. 2010. Autobi-a tool for automatic tobi annotation. In *INTERSPEECH*, pages 146–149.

Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Gaze patterns and fillers

Empirical data on the difference between Dutch 'euh' and 'euhm'

Annelies Jehoul
KU Leuven
annelies.jehoul@kuleuven.be

Geert Brône
KU Leuven
geert.brone@kuleuven.be

Kurt Feyaerts
KU Leuven
kurt.feyaerts@kuleuven.be

Abstract

In naturally occurring conversation, speakers use fillers such as 'euh' and 'euhm' for a variety of reasons. In this study, we explore speakers' gaze behavior when producing a filler, as the functions that have been associated with fillers and gaze aversion show some promising similarities. Studies show that both fillers and gaze aversion are associated with a speaker being hesitant or uncertain (Kendon, 1967, De Leeuw, 2007). Also in terms of the turn taking system, the functions of fillers and gaze aversion overlap as they both play an important role as turn holding signals (Maclay and Osgood, 1959; Kendon, 1967). However, the present analysis shows that speakers' gaze behavior when uttering a filler is not so clear cut: a sustained gaze at the interlocutor during a filler is almost as frequent as gazing away. In the second part of the analysis, we compare speakers' gaze patterns when producing two different manifestations of the filler ('euh' vs. 'euhm' in Dutch). These formal variants show some interesting differences in the co-occurring gaze distribution. More than 'euh', the longer variant 'euhm' is accompanied by a gaze aversion or a gaze fixation to the background. The multimodal analysis we present in this study supports previous findings of an interactional difference in vocal and vocal-nasalic fillers (Clark, 1994; Swerts, 1998; De Leeuw, 2007, Navarretta, 2015).

1 Introduction

In spontaneous conversation, speakers frequently produce verbal fillers such as 'euh' and 'euhm'. Various functions have been attributed to these small, seemingly meaningless elements. Clark and Fox Tree (2002) distinguish three views on the English 'uh' and 'um'. Traditionally, fillers are believed to be a symptom of production difficulties. According to this *filler-as-symptom* view, speakers use fillers when they are in doubt, for example when they have to make a choice, voice a challenging thought or have other difficulties planning their utterance. Fillers show the speaker's hesitation and uncertainty (De Leeuw, 2007; Clark and Fox Tree, 2002). Proof for this hypothesis was found in the fact that fillers are used more frequently in describing ambiguous occurrences and abstract nouns, elements that typically demand a more complex cognitive process (De Leeuw, 2007; Reynolds and Paivio, 1968; Siegman and Pope, 1966).

However, in the last decades, researchers found that this symptom hypothesis cannot be the only factor in accounting for the use of fillers. Lake, Humphreys and Cardy (2011) found that individuals with autism use fewer filled-pause words than a control group, and show in contrast more silent pauses and repetitions. Moreover, 'um' occurs less frequently in deceptive speech and in human-robot interaction than in natural human-human conversation (Villar, Arciuli and Mallard, 2012; Walter, Risko and Kingstone, 2014). These findings suggest that the use of fillers lies partly within a speaker's control. Following on this, fillers can be seen as a signal, or even a word. In the *filler-as-turn management-signal* view[1], fillers are considered as turn management signals that speakers give to their interlocutors. According to this view, a speaker can use a filler as a signal to hold, yield or take the turn. Speakers use

[1] This term corresponds to *filler-as-nonlinguistic-signal* as described by Clark & Fox Tree (2002). We chose to change the term to *filler-as-turn management-signal*, because the turn management function that fillers can fulfill is also a linguistic function.

fillers when they want to keep the floor, to indicate to their interlocutors that they are still in control of the turn (Maclay and Osgood, 1959; Clark and Fox Tree, 2002). In other contexts however, fillers can signal the speaker's willingness to give up the turn. They can, for example, ask for the interlocutor's co-participation in a speaker's word search sequence (Goodwin and Goodwin, 1986; De Leeuw, 2007). Fillers can also function as turn taking signals, when a speaker starts a turn with a filler just after the completion of the previous, to indicate that he wants the next turn (Beattie, 1983; Sacks cited in Schegloff, 1982). A third view on fillers is the *filler-as-word* view, in which fillers are considered as words, related to interjections, which announce a delay in speaking. According to this view, announcing a delay is a filler's primary function, while "most other functions are implicatures that follow from the relevance of announcing minor or major expected delays in the current situation" (Clark and Fox Tree, 2002). This third view, however, is criticized by O'Connell and Kowal (2005), who state that fillers do not function in the same way as interjections.

These different views show that, from the speaker's perspective, fillers can fulfil a communicative goal. Additionally, fillers also have an effect on the interlocutor. From the listener's perspective, fillers facilitate recall. When an interlocutor hears a filler, he is quicker to recognize the upcoming word. Also, a listener is more likely to anticipate on a less accessible referent when the word is preceded by a filler (Fraundorf and Watson, 2011; Brennan and Williams, 1995). These views aren't mutually exclusive though: when 'uh' stems from a production difficulty, for example when a speaker is searching for a word, the filler can also be used as a turn holding signal, and even announce a delay to the interlocutor. As fillers can express different functions, the context in which the filler appears plays an important role (Clark and Fox Tree, 2002).

This study explores the correlation between fillers and gaze aversion by a speaker. Gaze aversion, or breaking off the mutual gaze between participants (infra), arises for different communicative purposes, which seem to relate to certain functions of fillers. It is known that speakers look away to avoid an overload of possibly distracting visual information (Schober et al., 2012). Gazing away is therefore a frequent strategy in situations in which the speaker is planning an utterance, to focus on the cognitive resources, rather than on distracting information (Kendon, 1967; Weiß and Auer, 2016). As fillers can function as symptoms of production difficulties, it is likely that they correlate with gaze aversion, so the speaker can avoid further distraction during the production difficulties. Gaze aversion also fits in with the turn management functions of fillers, in which fillers are seen as turn managing cues. Speakers gaze away when they want to keep the turn, as gazing at the interlocutor can offer the latter the opportunity to take the turn (Argyle and Cook, 1976; Mutlu et al., 2012). This goal of gaze aversion fits in with the filler function of turn holding and a co-occurrence is therefore very likely. Gaze aversion also occurs when a speaker wants to take the turn: speakers avert their gaze to concentrate on planning the utterance (Argyle and Cook, 1976). Because of these similarities in communicative goals, fillers are expected to frequently co-occur with gaze aversion.

In Dutch, speakers produce two different manifestations of the filler, 'euh' and 'euhm', parallel to English 'uh' and 'um'.[2] As already observed before, the formal realization of a filler affects its communicative function. The vocal 'uh' seems to warn the interlocutor for a short interruption, whereas the vocalic-nasal 'um' announces a longer delay (Clark, 1994). Accordingly, vocalic-nasal fillers in American English are found more during planning of larger units, whereas a vocal filler reflects "local lexical decision-making" (Shriberg, 1994). A related finding in Dutch is that 'euhm' occurs more frequently at major discourse boundaries than 'euh' (Swerts, 1998; De Leeuw, 2007). This functional difference correlates with another difference in form: 'euhm' is more likely to be surrounded by pauses than 'euh' (De Leeuw, 2007). Also in Danish, the vocalic-nasal filler 'øhm' occurs more frequently preceding phrases than the purely vocal filler 'øh' (Navarretta, 2015). This study hypothesizes that this formal and functional difference between 'euh' and 'euhm' is reflected in a difference in gaze pattern.

2 Methodology

The distribution of the fillers 'euh' and 'euhm' was studied in three triadic interactions. The participants held a natural conversation of approximately 15 minutes. Some potential topics were suggested to the

[2] These two manifestations are found in Dutch, German and English, but their distribution differs. A comparative study of fillers in Dutch, English and German can be found in De Leeuw 2007.

groups, such as holiday plans, exchange program experiences, hobbies and social media. However, they were free to talk about anything they wanted, so they did not have to restrict themselves to the suggested topics. The participants were all acquainted students between 18 and 24 years old. Two of the groups contained two females and one male, one group consisted of only females. All participants were equipped with 'Pupil Pro Eye-Tracking Glasses' to track their eye movements.[3] The screenshot in figure 1 shows the recording setup of one of the interactions. Both upper and the left lower boxes show the dynamic eye tracker viewpoint, each filmed by one of the participants' eye tracking glasses. The lower right box shows the scene camera perspective on the interaction. The four camera perspectives were synchronized into one video file, resulting in a quadvid that shows the perspectives simultaneously. Figure 1 is a still from the synchronized video file.

Figure 1. Recording setup and resulting quadvid for the triadic interactions.

A transcription of the speech was made in the annotation tool ELAN (Wittenburg et al., 2006), in combination with the annotation of the gaze target during, before and after the filler, gestures made during, before and after the filler, the presence of empty pauses before or after the filler and the position of the filler in the turn.

3 Analysis

3.1 Fillers and gaze aversion

Before coming to the actual analysis of the co-occurrence of fillers and gaze aversion, a note should be made on the concept of gaze aversion. Gaze aversion is in this study characterized as a 'looking away', and can be realized in different ways, depending on the original area of interest and the area of interest the speaker switches to. A key characteristic is that the speaker shifts their gaze to a point that is less central than the previous point of focus. The gaze shift that shows gaze aversion most clearly is a shift from an interlocutor to the background, as this is a straightforward way to avoid interference from another speaker. Secondly, a speaker can avert their gaze with a shift from one focus in the background to another one, if the shift reflects a further aversion. For example, a speaker can shift their gaze from the wall, next to one of the interlocutors, to a point on the floor or the ceiling. Thirdly, a gaze aversion can be achieved by a shift between two interlocutors. In this way, the speaker interferes with the first target's possibilities to interrupt him. Opposed to this "looking away", speakers can perform a 'looking back'

[3] The eye gaze of one of the participants could not be studied, due to technical difficulties. The analyses therefore encompasses the gaze behavior of eight participants.

shift, during which they shift their gaze back to a more central point in the interaction. An example is a shift from the background to an interlocutor. A third possible gaze pattern is the sustained gaze at the background or an interlocutor, when a gaze shift does not occur in close temporal relation to a filler. Because of the functional correlation between fillers and gaze aversion, we expect a filler to be accompanied by gaze aversion, a 'looking away' more frequently than by a shift in the opposite direction ('looking back'). Also, when a speaker keeps their gaze fixated on one point during the production of the filler, this point is more likely to be an element in the background rather than an interlocutor.

A second note focuses on the time range within which the gaze shifts may occur. In this study, the speaker's gaze during the filler, during the pause that often precedes or follows a filler as well as during 500 ms before the onset of the filler and preceding pause is taken into account. A qualitative study of gaze shift surrounding fillers shows that this time frame captures the most gaze shifts related to the hesitation underlying the filler, and leaves out most shifts related to other elements in the interaction.

In the corpus, 81 fillers were found. Due to technical issues with the calibration of the eye-tracker, the gaze target of the speaker could only be reliably observed for 73 fillers. Analysis of these fillers shows that the expected pattern, a looking away just before or during a filler, is frequent: in 38% of the cases (28 fillers), the speakers avert their gaze. The opposite pattern, a looking back, is far less frequent. Only 9 of the 73 fillers are accompanied by a gaze that shifts back to a more central point in the interaction. 3 of the fillers are accompanied by more complex shifts, e.g. a shift from the background, to an interlocutor and back to the background. An unexpected result, however, is how frequently a speaker keeps their gaze fixated on an interlocutor during or just before fillers. This is a pattern that occurs in 27 instances (37%), whereas a fixated gaze on the background is far less frequent (6 fillers, 8%).[4]

gaze behavior	**frequency of fillers**	**relative frequency**
looking away	28	38%
looking back	9	12%
other shifts	3	4%
background	6	8%
interlocutor	27	37%
total	73	100%

Table 1. Gaze aversion during fillers.

These to some extend unexpected results can be explained by the diverse functions fillers can fulfill. As already described in section 2, fillers can be used for different purposes. Often, a filler expresses a hesitation or a planning pause by the speaker, but it can also be a sign that a speaker is polite, it can mark a syntactic boundary, etc. (Clark and Fox Tree, 2002). It is very likely that these different functions of fillers correlate with different gaze patterns, and that only fillers that truly express hesitation, co-occur with gaze aversion. A second, but related explanation may be found in the verbal form of the filler: a vocal filler could be accompanied by a different gaze pattern than a vocalic-nasal filler. This theory will be tested in the second part of this paper.

3.2 'euh' vs. 'euhm'

Next to a general correlation between fillers and gaze aversion, this study analyzes the difference in gaze pattern during the different fillers 'euh' and 'euhm'. In the three studied triads, the vocal filler 'euh' occurs far more frequently than the vocalic-nasal 'euhm'. A total of 66 instances of euh were found, whereas 'euhm' only occurs 15 times in the corpus. This low frequency implies that remarks about the difference between the two manifestations of the filler should be made with caution.

[4] Although there are some minor individual differences, the gaze behavior of six of the eight participants conforms to this pattern.

As table 2 shows, there is a substantial and significant difference in gaze pattern between 'euh' and 'euhm' (Fisher's exact test, $p < 0.01$). The most striking difference lies in the number of cases of gaze remaining fixated on the interlocutor (i.e. the category with the somewhat unexpected outcome in the first section of the analysis. In this more detailed analysis, we can see that almost all of the fillers during which the speaker keeps gazing continuously at an interlocutor, are instances of 'euh'. It happens only once that a speaker keeps gazing at an interlocutor when uttering 'euhm'.

gaze pattern	***euh***		***euhm***	
	absolute frequency	**relative frequency**	**absolute frequency**	**relative frequency**
looking away	20	34%	8	53%
looking back	4	7%	5	33%
other shifts	2	3%	1	7%
background	6	10%	0	0%
interlocutor	26	45%	1	7%
total	58	100%	15	100%

Table 2. Formal variants of filler and gaze aversion.

This difference in co-occurrence between gaze behavior and different kinds of fillers corresponds with the functional difference described in section 2 of this paper. All formal and functional differences show the same tendency of the vocalic-nasal filler 'euhm' being the signal for a deeper cognitive thinking process. A vocalic-nasal filler ('euhm' in Dutch) is more frequent in the planning of larger units, occurs more frequently at major discourse boundaries and announces a longer delay than the purely vocalic counterpart ('euh' in Dutch) (Shriberg, 1994; Swerts, 1998; Clark, 1994). This functional difference is, as the results of this exploratory analysis show, also reflected in the speakers' gaze behavior. When speakers avert their gaze more frequently when producing 'euhm', this may very well be related to the deeper cognitive process that the filler signals. This function is shown in example 1. In this example, one of the speakers, Sharon, asks Veronique how many exams she has to take. Veronique answers this question, but she starts her answer with an 'euhm' and a longer pause. Veronique already averts her gaze before Sharon's question is finished. This gaze aversion, together with the 'euhm' functions as a display of a cognitive process: Veronique is probably counting her exams in order to be able to provide a correct answer. She looks back at Sharon, who asked the question, when she has almost finished her answer ('vier', *four*).

```
Example 1
1   SHARON      en hoeveel examens hebde gij?
                and how many exams do you have?
    gaze VER    SHARON                  BACKGR.

2   VERONIQUE   euhm:: (-) vier (-)
                um::   (-) four (-)
    gaze VER    BACKGROUND      SHARON

3   VERONIQUE   ja
                yes
    gaze VER    SHA.

4   VERONIQUE   en gij?
                and you?
    gaze VER    SHARON
```

This is a typical situation in which the speakers in the studied triads use ‘euhm’. The pure vocalic filler ‘euh’, however, is used in far more varied situations. An example can be seen in example 2. Maarten is talking about a Spanish girl who he and Sharon met during their stay in Spain. He is searching for the name of one of their friends who lived with the Spanish girl. During this word search, he utters ‘euh’ two times. While he is searching for the name, he keeps gazing at Shana, apparently to appeal to her to finish his sentence and say the name of the friend. In this case, the filler ‘euh’ is not used as a turn holding cue, but rather as a turn yielding cue. This example supports Goodwin and Goodwin’s (1986) findings that during a word search, when a speaker’s gaze is directed towards an interlocutor, the speaker asks the interlocutor to participate in the word search. In this word search, this gaze behavior is successful: Sharon finishes Maarten’s utterance and supplies the name he’s searching for.

Example 2

1	MAARTEN	ma ja da was wel goeie kotgenoten twas geen vieze
		but yeah that was good roommates it was no dirty
	gaze MAA	SHARON
		mexicaan of
		mexican or
	gaze MAA	SHARON
2	SHARON	h[a]
		h[a]
	MAARTEN	[of] die hysterische spaanse van: euh: (.)
		*[or] that hysterical spanish girl from **uh:**(.)*
	gaze MAA	SHARON
3	MAARTEN	in: euh:
		in: uh:
	gaze MAA	SHARON
4	SHARON	ja [oh:] Eva
		yeah [oh:] Eva
	MAARTEN	[lacht]
		[laughing]
	gaze MAA	SHARON NO DATA

4 Discussion

The first hypothesis in this study, stating that fillers are often combined with gaze aversion, could not be supported. People do avert their gaze during or just before 38% of the fillers, but they also keep gazing at their interlocutor during 37% of the fillers. Other patterns, such as a constant gaze at the background or a gaze shift back to the interlocutor are less frequent. However, the data do support the second hypothesis as a correlation between the kind of filler and gaze patterns could be found: speakers in the corpus avert their gaze more often during ‘euhm’ than during ‘euh’. The sustained gaze at the speaker, which turned out to be more frequent than expected in the first part of the analysis, only occurs once during ‘euhm’, but is a very persistent pattern for ‘euh’. These divergent gaze patterns can be attributed to the functional difference between the vocal and vocalic-nasal fillers, a tendency that occurs in most Germanic languages. The nasal variant ‘euhm’ occurs frequently at major discourse boundaries, reflecting a cognitive thinking process, typically when the speaker is planning their discourse. ‘Euh’, on the contrary, is used more often when speakers are involved in a word search, making a more local lexical decision (Swerts, 1998; Shriberg, 1994; Navarretta, 2015). Since this study only analyzed a limited

number of fillers, a more comprehensive corpus study, analyzing more attestations of both kinds of fillers, is needed to strengthen these results.

The different gaze pattern during vocal and vocalic-nasal fillers sheds some light on the gaze distribution during fillers, but cannot account for all of the differences in gaze distribution. Especially during occurrences of 'euh', we have observed quite a lot of variation in the gaze pattern. Qualitative research on the function of these fillers should point out whether this differentiated gaze behavior corresponds with different functions. In this respect, also other features of fillers, such as the duration of the filler and contextual gesture should be taken into account. Since the gaze behavior in 'euh' and 'euhm' diverges considerably, a correlation between gaze behavior and functional differences between fillers is likely to occur.

References

Michael Argyle and Mark Cook. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, UK.

Geoffrey W. Beattie. 1983. *Talk: An Analysis of Speech and Non-verbal Behaviour in Conversation*. Open University Press, Milton Keynes, UK.

Susan E. Brennan and Maurice Williams. 1995. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34: 383-398.

Herbert H. Clark. 1994. Managing problems in speaking. *Speech Communication*, 15(3-4): 243-250.

Herbert H. Clark and Joan E. Fox Tree. 2002. Using 'uh' and 'um' in spontaneous speaking. *Cognition*, 84: 73-111.

Esther De Leeuw. 2007. Hesitation markers in English, German, and Dutch. *Journal of German Linguistics*, 19(2): 85-114.

Scott H. Fraundorf and Duane G. Watson. 2014. Alice's adventures in umderland: psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, 29(9): 1083-1096.

Marjorie H. Goodwin and Charles Goodwin. 1986. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62: 51-75.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26: 22-63.

Johanna K. Lake, Karin R. Humphreys and Shannon Cardy. 2011. Listener vs. speaker-oriented aspects of speech: studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review*, 18(1): 135-140.

Howard Maclay and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15: 19-44.

Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Trans. Interact. Intell. Syst.*, 1(2): article 12.

Costanza Navarretta. 2015. The functions of fillers, filled pauses and co-occurring gestures in Danish dyadic conversations. *Proceedings from the 3rd European Symposium on Multimodal Communication*, 55-61.

Daniel C. O'Connell and Sabine Kowal. 2005. 'Uh' and 'um' revisited: are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34(6): 555-576.

Allan Reynolds and Allan Paivio. 1968. Cognitive and emotional determinants of speech. *Canadian Journal of Psychology*, 22: 164–175.

Elizabeth Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California, Berkeley.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: some uses of "uh huh" and other things thatcome between sentences. In Deborah Tannen (Ed.), *Analyzing discourse: text and talk*. Georgetown University, Washington, DC.

Michael F. Schober, Frederick G. Conrad, Wil Dijkstra and Yfke P. Ongena. 2012. Disfluencies and gaze aversion in unreliable responses to survey questions. *Journal of Official Statistics*, 28(4): 555-582.

Aron W. Siegman and Benjamin Pope. 1966. Ambiguity and verbal fluency in the TAT. *Journal of Consulting Psychology*, 30: 239-245.

Marc Swerts. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30: 485-496.

Gina Villar, Joanne Arciuli and David Mallard. 2012. Use of "um" in the deceptive speech of a convicted murderer. *Applied Psycholinguistics*, 33: 83-95.

Esther J. Walker, Evan F. Risko and Alan Kingstone. 2014. Fillers as signals: evidence from a question – answering paradigm. *Discourse Processes*, 51(3): 264-286.

Clarissa Weiß and Peter Auer. 2016. Das Blickverhalten des Rezipienten bei Sprecherhäsitionen: eine explorative Studie. *Gesprächsforschung*, 17: 132-167.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*.

Gestures become more informative after negative feedback

Marieke Hoetjes
Radboud University
Centre for Language Studies
Nijmegen, The Netherlands
m.hoetjes@let.ru.nl

Abstract

When objects or people are described repeatedly in conversation, the repeated references tend to be reduced variants of initial references. This can be seen both in speech, and in gesture. Previous studies focused on successful repeated references, produced in contexts of common ground. A question is whether repeated references are also reduced in contexts where there is less, or no common ground, for example during communicative problems. In particular, the present study asks whether gestures which are produced in repeated references following negative feedback become more informative for the addressee. Participants viewed silent video clips, each showing one gesture, taken either from object descriptions before any feedback was given, or from object descriptions given after (repeated) negative feedback. With each video clip participants were shown two objects. The task was to decide which of the two objects was the target associated with the gesture they were shown. Results showed that participants were better at this task when presented with gestures produced following (repeated) negative feedback. This leads us to conclude, firstly, that after having received negative feedback, gestures are not reduced, but become more informative, and secondly, that this might be done with the addressee in mind.

1 Introduction

When people communicate, they often refer to particular objects or people. For example, in a conversation about pets, someone might mention "that small ginger cat". This referring expression may be multimodal, that is, the speaker may accompany speech with hand gestures, for example one gesture indicating the size and one gesture indicating the location of the cat in question. Also, the same cat might be referred to more than once during the conversation. Previous studies have shown that when people produce such repeated references, these repeated references are often reduced, lexically, acoustically, and gesturally. In a seminal study by Clark and Wilkes-Gibbs (1986), participants had to repeatedly describe the same tangram figures. It was found that repeated descriptions were lexically reduced, for example from an initial description of "a person who's ice skating, except they're sticking two arms out front", to a sixth description of the same figure as "the ice skater" (Clark & Wilkes-Gibbs, 1986, p. 12). In the case of our pet example, a repeated reference to the cat that is lexically reduced could be "the cat". Repeated references have also been shown to be reduced acoustically. Bard et al. (2000), for example, found that references to given information were less intelligible when they were taken out of context and presented to naïve listeners. Finally, repeated references have also been found to be reduced with regard to gesture. Previous studies (e.g. Galati & Brennan, 2014; Jacobs & Garnham, 2007) found that the number of gestures is lower in repeated references than in initial descriptions. This reduction in gesture in repeated references is not that surprising, given that speech and gesture are closely related (Kendon, 2004; McNeill, 1992) and tend to be co-expressive.

The reduction process in repeated references can be explained by the fact that repeated references are usually produced in a context of common ground (Clark & Brennan, 1991). After all, after an ini-

tial description has introduced the object, there is common ground between interlocutors, and a reduced repeated description is sufficient to still know which object is being discussed. In line with this, previous work (Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015) showed that gestures in repeated references were reduced with regard to their number and precision. However, this reduction did not make the gestures less informative. In a gesture interpretation experiment, Hoetjes, Koolen, et al. (2015) presented the less precise gestures, produced in repeated references, to addressees, and it was found that addressees were equally likely to link them to the object they referred to as they did for the gestures from initial references, which were not reduced.

Studies like the ones mentioned above studied repeated references in contexts in which communication was successful, either because interlocutors gave each other explicit positive feedback, or implicit positive feedback (e.g. because the correct object was selected). In these cases there is common ground between interlocutors, and repeated references can easily be reduced without causing communication problems. A question is whether such reduction processes in repeated references also occur in cases of communicative problems, when arguably there is less, or no, common ground between speakers. If there are communicative problems, these may become apparent because one of the interlocutors gives negative feedback (e.g. by saying "Sorry, which cat", or by not identifying the cat in question). Presumably, a following repeated reference to the same object may not be reduced but may instead be enhanced somehow, for example by not reducing the number of words or gestures but by keeping them constant, or even increasing them (e.g. negative feedback could cause a repeated reference to the above mentioned cat to become "the small ginger cat over there with a stripy tail"), so that the addressee is more likely to correctly identify the target object.

Focusing on gesture production, only a few studies have been done that address the question whether there is also reduction in repeated references in cases of communicative problems. Holler and Wilkin (2011) conducted a study in which participants had to retell fragments from a television series to a confederate addressee. The addressee gave scripted negative feedback (e.g. requesting clarification) at predetermined points in the narrative. Because of this feedback, participants were required to re-describe part of their retellings. When comparing 100 pairs of gestures that were produced before and after the feedback, it was found that in 60% of the cases, the gestures became either larger, more precise, or visually more prominent. Holler and Wilkin state that this change in gesture production means that utterances became clearer for the addressee. In other words, the negative feedback led to gestures that were more informative for the addressee than the gestures produced before the negative feedback.

In a study by Hoetjes, Krahmer and Swerts (2015) participants had to describe objects to a confederate addressee, who had to identify the target object from a set of objects. In several cases, the addressee provided negative feedback by identifying the incorrect object. This negative feedback meant that the participant had to describe the same object again, until it had been correctly identified. There were several objects that each participant had to describe three times immediately after another (two of these descriptions occurred after negative feedback). In this production experiment, it was found that the repeated descriptions produced after negative feedback were reduced with regard to the number of words but not the number of gestures, causing an increase in relative gesture rate. Moreover, in line with the study by Holler and Wilkin (2011), a separate perception experiment showed that the gestures produced in repeated references after negative feedback were considered marginally more precise than the gestures produced before any feedback was given.

The results from these two studies show that, unlike in the studies where references were repeated in contexts of common ground, after receiving negative feedback, gestures in repeated references are not reduced. This means that repeated references are not always reduced variants of initial references, but whether the object description is reduced depends on the communicative context (i.e. whether there is common ground or not). Although these findings are in themselves interesting in relation to previous work on repeated references, we do not yet know whether the increased accuracy in gesture after negative feedback is also communicatively meaningful. That is, does the fact that gestures in repeated references after negative feedback become more precise also help the addressee in identifying the correct object? Therefore, the question addressed in the present study is whether gestures that are produced following negative feedback become more informative for an addressee. If this is the case, we propose that this change in gesture production by the speaker could be done with the addressee in mind. This study builds upon work conducted in the previously mentioned studies by Hoetjes, Krah-

mer and Swerts (2015) and by Hoetjes, Koolen, et al. (2015). Specifically, it uses the same material as in the perception experiment by Hoetjes, Krahmer and Swerts (2015), and the same procedure as in the gesture interpretation experiment in Hoetjes, Koolen, et al. (2015).

2 Method

2.1 Participants

Sixty-nine participants (21 males, *M* = 21 years old, range 18-28 years old) took part in this study. The participants were undergraduate students who received partial course credits. None of the participants had taken part in any of our previous studies on gestures in repeated references (as reported in Hoetjes, Koolen, et al., 2015; Hoetjes, Krahmer, et al., 2015).

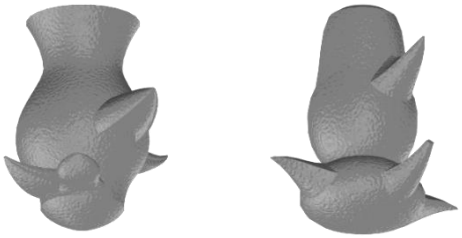
2.2 Material

Participants were presented with 88 short video clips. The video clips were played without sound, to avoid any influence of speech. Each video clip showed someone producing one gesture, and lasted between 1 and 6 seconds. The video clips were the exact same as used in the perception judgment experiment in our previous work on repeated references (Hoetjes, Krahmer, et al., 2015), which were taken from the recordings of that study's production experiment. The 88 video clips, consisting of 44 pairs of video clips, were selected as follows. Video pairs consisted of one video showing a gesture from an initial description, and one video showing a gesture produced after negative feedback (so produced during a second or third reference). The video pairs showed gestures produced by the same speaker, and both gestures referred to the same part of the same object. To avoid overrepresentation of a small number of speakers, no more than two gesture pairs from each speaker were used. Pairs of video clips always consisted of one video taken from an initial reference, and one video taken from a repeated (second or third) reference. There was a fairly equal distribution between gestures from second and from third references (23 of the pairs were taken from initial and second descriptions, and 21 pairs were taken from initial and third descriptions). The gestures in the video clips were all iconic gestures, illustrating an aspect of the object that was being described. In total, this lead to a set of 44 gesture video clip pairs – 88 video clips in total. 44 video clips showed a gesture produced in an initial reference (before feedback), 22 videos showed a gesture produced in a second reference (after negative feedback), and 22 videos showed a gesture produced in a third reference (after repeated negative feedback). The video clips were presented individually (not in pairs), and semi-randomly, such that two video clips of one pair were never presented one after another. A still of one of the video clips can be seen in figure 1.

2.3 Instruments

For each video clip, the participants were presented with two pictures (A and B) on one piece of paper. These pairs of pictures always showed one object that was actually being described in the video clip (i.e. the correct object), and one object which looked similar but which had a main 'body' shape which was different from the other object (i.e. the incorrect object). The order in which the correct object was being presented (A or B) was counterbalanced over trials. An example picture set showing the answer possibilities for one trial can be seen in figure 2.

Figure 1. Still from one of the video clips. The arrows indicate path and direction of the gesture.



A B

Figure 2. Example of answer possibilities.

2.4 Procedure

Participants took part in the experiment individually. The video clips were presented one after another in a PowerPoint presentation, and participants were free to go through this presentation of the 88 video clips by themselves. Video clips started playing as soon as a new slide was opened and participants were only allowed to see each video clip once. For each video clip there was a separate piece of paper with the two answer possibilities for that particular trial on it (as in Figure 2). The task was to decide, on the basis of the one gesture shown in the video clip, whether the gesture was produced during the description of object A or object B. Participants filled in their answer on an answer form. Before the experiment started, participants were given written instructions, the opportunity to ask questions, and two practice trials to help them get used to the short video clips. The entire experiment took about 25 minutes.

2.5 Design

The experiment had one independent variable, repetition, with three levels (initial, second, third). The study was set up in a within subject design. Each participant was presented with all video clips. The video clips showed gestures that were produced during initial references (preceding feedback), second references (following negative feedback), or third references (following repeated negative feedback).

3 Results

The number of times that participants chose the correct object was counted and analysed using a chi-square analysis. The results can be found below in table 1. We found that there was a significant association between repetition and the number of times that the correct object was selected, $\chi^2(2) = 23.290$, $p < .001$. If we look at the distribution of percentage of correct versus incorrect answers, we see that for gestures produced during initial and third references, there were more correct than incorrect answers (as indicated by the subscripts in Table 1). If we look at the percentage of correct answers across the three conditions, we can see that there were more correct answers (54.8%) for gestures produced after initial negative feedback (second references) than for gestures produced before any feedback was given (53%), and even more correct answers (60.5%) for gestures produced after repeated negative feedback (third references).

Table 1. Number (and percentages) of correct and incorrect answers, across conditions (initial, second and third references, 2nd and 3rd references were produced after negative feedback). Within conditions, subscripts indicate significant differences between percentage of correct and incorrect answers.

	Initial	Second	Third	Total
Correct	1608_a *(53%)*	832_a *(54.8%)*	918_a *(60.5%)*	3358 *(55.3%)*
Incorrect	1428_b *(47%)*	686_a *(45.2%)*	600_b *(39.5%)*	2714 *(44.7%)*
Total	3036 *(100%)*	1518 *(100%)*	1518 *(100%)*	6072 *(100%)*

4 Discussion and Conclusion

The research question of this study was whether negative feedback would change gesture production in repeated references in such a way that gestures would become more informative for a naïve viewer. The results showed that as more negative feedback was given (especially in third references), participants more often correctly selected the object during which description the gesture was originally produced. We can therefore conclude that gestures after negative feedback become more informative.

The findings complement previous work on gesture production in repeated references after negative feedback. In particular, the studies by Holler and Wilkin (2011) and by Hoetjes, Krahmer and Swerts (2015) showed that when interlocutors provide feedback that indicates that there was some sort of communicative problem (e.g. by explicitly asking for more information, or by selecting the incorrect referent), gestures in repeated references are not reduced, but can increase, with regard to their size, precision, or prominence. However, it was previously unclear whether these changes in gesture production are also useful for an addressee. We can now provide evidence that the changes in gesture

production that occur in repeated references when communication is unsuccessful are in fact useful for an addressee and might be done with this addressee in mind.

We can relate the findings of this study also to previous work on gesture production in repeated references where there was no negative feedback. Specifically the study by Hoetjes, Koolen, et al. (2015) found that in repeated references, speakers produced fewer and less precise gestures. When conducting their gesture interpretation experiment however, it turned out that these changes in gesture production did not make the gestures less informative, i.e. in their experiment they found that participants were equally likely to correctly select the target object based on a gesture from an initial or from a repeated reference. In the current study, the same procedure was used. We can therefore directly compare the results of their gesture interpretation experiment to the results of the current study. In the study by Hoetjes, Koolen, et al. (2015) object descriptions were repeated in a context of common ground, without communicative problems. In the current study the object descriptions were repeated because the addressee provided negative feedback, indicating that there were communicative problems. Combining the findings from both studies it can be concluded that when gestures are produced during repeated object descriptions, they only become more informative if the discourse context requires it. When there are no communicative problems and there is common ground between speaker and addressee, there is no need to make the gesture more informative for the addressee. When negative feedback indicates that there are communicative problems, and consequently there is less, or no, common ground, gesture production in repeated references is adapted in such a way that the gesture can help the addressee in correctly identifying the target object.

To conclude, this study suggests that gestures can provide valuable information in a discourse context. In this case, participants were able to select the correct object above chance level, after only viewing one gesture (which is a hard task, especially without the original speech), and this ability increased when these gestures were produced after negative feedback. Based on these findings, we would like to claim that by adapting their gestures when communication is unsuccessful in such a way that they become more informative, speakers help the addressee, and thereby help to keep the overall communicative situation as successful as possible.

Acknowledgments

Many thanks to Malu Hanssen for her help in collecting the data.

References

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*, 1-22.

Clark, H., & Brennan, S. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & J. S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149): American Psychological Association.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Galati, A., & Brennan, S. (2014). Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience, 29*(4), 435-451. doi: 10.1080/01690965.2013.796397

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language, 79-80*, 1-17.

Hoetjes, M., Krahmer, E., & Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Communication, 72*, 160-175.

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics, 43*, 3522–3536.

Jacobs, N., & Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language, 56*, 291-303.

Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge: Cambridge University Press.

McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. Chicago: University of Chicago Press.

Can gestures change perceived meaning of ambiguous motion events Evidence from Italian verb-particle constructions

Bjørn Wessel-Tolvig
University of Copenhagen, CST
bwt@hum.ku.dk

Patrizia Paggio
University of Copenhagen, CST
University of Malta, ILLT
paggio@hum.ku.dk

Abstract

How sensitive are Italian speakers to information provided by co-speech gestures when interpreting ambiguous motion events? Verb + particle constructions are not suitable for expressing telic motion (change of location across a spatial boundary) in verb-framed languages like Italian. However, this constraint may perhaps be disregarded with certain type of manner verbs + complex PPs. The reading often depends on contextual inference or pragmatic clues. We present two experimental judgment tasks in which we first test whether grammatically locative Italian verb + particle constructions can be interpreted as boundary-crossing motion and secondly we investigate the effect of gestural information on the same type of locative events. The study confirms the existence of boundary-crossing interpretations for certain types of Italian manner verb + PP constructions, but more importantly that co-speech gestures can change the reading of events and thus override default meaning expressed only in speech.

1 Introduction

Iconic gestures contain a lot of information about what we say and how we say it (Kita et al., 2007; Gullberg, 2011). These types of gestures are tightly linked to language and often reflect the same information as speech (McNeill, 2005). However, gestures may also express different aspects of that meaning (Beattie and Shovelton, 1999) and may therefore reveal more about what the speaker is trying to convey than speech alone (Athanasopoulos and Bylund, 2013). Co-speech gestures may thus help the listener gain information about speaker intentions and ideas especially in noisy environments (Harrison, 2011) or ambiguous situations (Goodrich Smith and Hudson Kam, 2012). Many studies investigating the integration of speech and gesture in comprehension look at situations where there is a mismatch or incongruence between what is expressed in speech and in gestures (Kelly et al., 2014; Holle and Gunter, 2007). However, in this paper we investigate the effects of gesture information on interpretation of information in a truly ambiguous areas of linguistics: the directional reading of locative particles (Folli, 2008; Gehrke, 2007). We set up two experimental judgment tasks to first test the interpretation of ambiguous manner verb + locative particles for expressing directional motion (NO GESTURE CONDITION), and secondly in a GESTURE CONDITION we test whether information in gestures has an effect on, and may alter, how the same motion constructions are interpreted.

2 Background

Recent research has indicated that the typology outlined by Talmy (1985; 1991) is too simple and rigid (Beavers et al., 2010). According to (Talmy, 2000), languages are generally classified in respect to how speakers of a particular language most typically express path of motion in lexical items within a clause structure. As is characteristic of a *verb-framed language* (Talmy, 1991), in Italian the path of motion (directionality) is typically expressed in the main verb of a clause and manner of motion, if present at all, is left to be expressed in a separate constituent (here an adverbial gerund) as in (1). However, the

typology refers to how motion is most frequently expressed and is not an absolute rule (Cadierno and Ruiz, 2006). Languages typically have other possibilities not typical of their framing type (Beavers et al., 2010; Croft et al., 2010), and Italian may frame the path of motion in ways typical of *satellite-framed languages* as in (2), where path is expressed in a satellite position to a main manner verb. Yet, verb-framed languages are constrained by the boundary-crossing constraint not to express telic motion across spatial boundaries with path expressed in satellites to main manner verbs (Slobin and Hoiting, 1994; Aske, 1989). The absence in Italian of inherently directional prepositions expressing change of location with end-goal and boundary-crossing properties (Iacobini and Vergaro, 2014), implies that motion across a spatial boundary cannot be constructed with a manner verb and a path-denoting satellite. In this case the satellite only denotes locative motion and the figure does not change states from outside to within the container (the goal), but only moves inside of the container as shown in (3).

1. Il pallone [FIGURE] **scende** [PATH] per la collina [GROUND] *rotolando* [MANNER]
   'The ball **descends** on the hill *rolling*'

2. Il pallone [FIGURE] **rotola** [MANNER] *giù* [PATH] per la collina [GROUND]
   'The ball **rolls** *down* the hill'

3. Il pallone [FIGURE] **rotola** [MANNER] *dentro* [LOCATION] la rete [GROUND]
   'The ball **rolls** **into/inside of* the goal'

Yet, recent developments in lexical semantics suggests that Italian speakers may overcome this constraint in two specific ways: either by combining certain types of manner of motion verbs with simple PPs, or possibly all types of manner verbs with complex PPs (Folli, 2008). According to Folli & Ramchand (2005, p. 97) some manner verbs in Italian carry an optional diacritic feature '+R' (=result), which licenses the projection of a result phrase (RP) specifying the end-goal of motion. Such verbs (e.g. *correre* - 'run') can in combination with a locative preposition be read as directional motion across a spatial boundary. On the contrary, pure manner verbs, which do not encode a result feature in their lexical specification (e.g. *danzare* - 'dance') only express locative motion when combined with a locative preposition. Folli (2008) extends the hypothesis to include all manner of motion verb types in combination with complex locative prepositions i.e., two or more prepositions (*dentro a, dietro a* - 'inside to/at', 'behind to/at') as in (4), and argues that two locative prepositions have a complex functional structure; one encodes a PATH/PROCESS component and the other a PLACE/END-LOCATION component. Thus the combination of a complex PP and a manner verb allows for a boundary-crossing reading of the event regardless of the verb type (Cardini, 2012), but the constructions are ambiguous in the sense that they also can denote locative meaning. The claim is contested by Mateu & Rigau (2010) and Bandecchi (2012) who both maintain that only manner verbs with a '+R' feature can be used to express directional motion with simple and complex PPs in Italian.

4. Il pallone [FIGURE] **rotola** [MANNER] *dentro alla* [PATH/LOCATION] rete [GROUND]
   'The ball **rolls** *into/inside of* the goal'

In fact, the construction is infrequently used in Italian (Wessel-Tolvig, 2015). The infrequency may be ascribed to Slobin's (1996) thinking-for-speaking hypothesis, which focuses on the potential effects language has on conceptualization. According to Slobin (1987; 2004) the language you speak, specifically the way manner and path are most frequently expressed in that language, has an effect on conceptualization in the process of interpreting and verbalizing motion events (Cadierno, 2012; Berman and Slobin, 1994). These thinking-for-speaking patterns may be so deeply rooted in cognition that *possible* manner verb + complex PP constructions may be biased in interpretation towards standard verb-framed locative meanings, i.e. a phrase like (4) is more likely to receive locative meaning than directional meaning. Since the manner verb + complex PP construction is ambiguous in expressing boundary-crossing motion, it may be difficult to infer speaker meanings based on speech alone. The co-expressive semantic content of gestures may provide listeners with an important additional indication of the speaker's intended meaning.

Speech and gesture are tightly related both semantically and temporally in language production (Kendon, 1980; McNeill, 1992). Studies show how speakers' co-speech gestures reflect what information they select for expression and how they express it (i.e. linguistic conceptualization) (Kita and Özyürek, 2003; Özyürek et al., 2005; Stam, 2006). Moreover, recent findings extend claims on the integration of speech and gesture to also hold for language comprehension (Kelly et al., 2010; Kelly et al., 2014; Holle and Gunter, 2007). Listeners incorporate information in speakers' gestures to derive speaker meanings (Dick et al., 2009) and thus attept to access speaker conceptualizations (Goodrich Smith and Hudson Kam, 2015). Under this perspective, the information in co-speech gestures may help, or guide, the listener when interpreting ambiguous expressions.

2.1 Research questions

Based on the recent proposal by Folli (2008) and Folli & Ramchand (2005) that Italian locative particles in combination with with complex PPs may be interpreted as expressing change of location and boundary-crossing movement, and given the tight relation between speech, gesture and cognition (Gullberg, 2011; McNeill, 2005), we ask the following questions:

- Can Italian complex locative PPs be assigned a boundary-crossing interpretation, and is the reading of such verb particle constructions influenced by lexical properties of the verb?
- Do listeners integrate information in co-speech gestures, and may gesture information influence the default interpretation of motion events?

3 Methodology

The data come from two independent experimental judgment tasks: a NO GESTURE CONDITION (baseline) and a GESTURE CONDITION involving a total of 212 participants, all native speakers of Italian. The judgment tasks are online questionnaires produced with Google Forms.

3.1 Participants

All participants are Italian native speakers recruited from all over the Italian peninsula and the data is collected using convenience and snowball sampling methods through personal and student networks (see table 1).

Table 1: Participant data

Condition	Participants	Gender (Female)	Age Mean(SD)
NO GESTURE	109	61.4%	27.6 (7.8)
GESTURE	103	69.9%	26 (6)

3.2 Experimental design

In the NO GESTURE CONDITION participants are shown different Italian manner verb + complex PP sentences in written form (see figure 1), and asked to judge if they understood the sentences as locative or directional motion (i.e. as movement within or into something). TYPE 1 verbs are manner verbs with a result feature encoded in the lexical specification, and TYPE 2 verbs are pure manner verbs that do not encode any result features. In the GESTURE CONDITION participants are asked to judge the same motion sentences as in the NO GESTURE CONDITION, however in a video-based format where the speaker produced either DIRECTIONAL or NON DIRECTIONAL gestures together with the sentences (see figure 2). Manner verbs belonging to the two verb types as defined in Folli & Ramchand (2005, p. 97) are included in both conditions. Verbs from the two groups are combined with the same complex PPs (e.g. *dentro a, fuori da* - 'into/inside of', 'out/outside of'), again in both conditions. Furthermore, in the GESTURE CONDITION half of the TYPE 1 verbs + PP are expressed with a DIRECTIONAL gesture and the other half with a NON DIRECTIONAL gesture (likewise for TYPE 2 verbs) as seen in table 2. All gesture strokes

are aligned with the main manner verb + PP + ground NP. The NO GESTURE CONDITION contains 12 motion event expressions and the GESTURE CONDITION 16. They are, however, equally distributed between verb types and PP types.

Figure 1: Example of elicitation material from the NO GESTURE CONDITION

Il pallone rotola dentro alla casa

('The ball rolls into/inside the house')

What did you first understand?

a) Il pallone entra nella casa
'The ball enters into the house'

b) Il pallone si trova già nella casa
'The ball is already in the house'

Figure 2: Example of video elicitation material from the GESTURE CONDITION

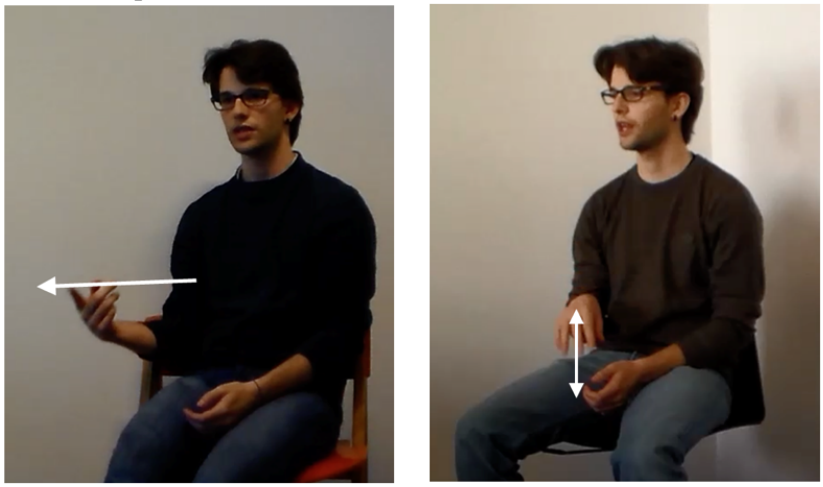


Table 2: Motion event construction and gesture type combination

Verb type	Verb (examples)	Complex PP	Gesture condition
TYPE 1	Rotolare (roll)	Dentro a (inside/into)	DIRECTIONAL
	Saltare (jump)	Fuori da (outside of/out of)	NON DIRECTIONAL
	Rimbalzare (bounce)	Dentro a (inside/into)	DIRECTIONAL
	Volare (fly)	Fuori da (outside of/out of)	NON DIRECTIONAL
TYPE 2	Galleggiare (float)	Dentro a (inside/into)	DIRECTIONAL
	Danzare (dance)	Fuori da (outside of/out of)	NON DIRECTIONAL
	Zoppicare (limp)	Dentro a (inside/into)	DIRECTIONAL
	Nuotare (swim)	Fuori da (outside of/out of)	NON DIRECTIONAL

4 Analysis

In the NO GESTURE CONDITION we collected a total of 1308 responses and in the GESTURE CONDITION we collected 1648.

4.1 NO GESTURE CONDITION

We find that the participants interpret TYPE 1 verbs with complex PPs as directional movement more often than they do TYPE 2 (see table 3). A one-way ANOVA with Verb type and Response frequency as within-group variables and Subject as error term, shows a main effect of Verb type $F(1,108) = 52$, $p < 0.001$, $\eta^2 = 0.0002$. Bonferroni-Holm corrected pairwise t-tests show significant differences in responses within each verb type. The finding provides evidence for the hypothesis that Italian manner verbs with an inherent directional result feature (TYPE 1) can give rise to boundary-crossing interpretations (Cardini, 2012; Folli and Ramchand, 2005). Yet, we also find that 20% of TYPE 2 verb + PP constructions are interpreted as movement across a spatial boundary even though this interpretation should not be available in theory (Bandecchi, 2012). This finding supports Folli's (2008) claim that the complex functional structure of the complex PP may give license to boundary-crossing interpretations even when combined with pure manner verbs (TYPE 2).

Table 3: Distribution of answers in the NO GESTURE CONDITION (absolute and relative frequencies)

	Manner verb type	
Interpretation	TYPE 1	TYPE 2
DIRECTIONAL	372 (.57)	134 (.20)
LOCATIVE	282 (.43)	520 (.80)
Sum	654 (1)	654 (1)

4.2 GESTURE CONDITION

In the GESTURE CONDITION we add the gesture variable as seen in the methodological section (table 2). Again we find that listeners are more likely to interpret the manner verb + complex PP constructions as boundary-crossing when the verb itself licenses some form of directional movement (TYPE 1) compared to pure manner verbs (TYPE 2) (see table 4). In a one-way ANOVA with Verb type as within-group variable and Subject as error term, we find a main effect of Verb type $F(1,102) = 5.47$, $p = 0.02$, $\eta^2 = 0.001$. Bonferroni-Holm corrected pairwise t-tests show significant differences in responses within each verb type.

Table 4: Distribution of answers in the GESTURE CONDITION (absolute and relative frequencies)

	Manner verb type	
Interpretation	TYPE 1	TYPE 2
DIRECTIONAL	441 (.54)	284 (.34)
LOCATIVE	383 (.46)	540 (.66)
Sum	824 (1)	824 (1)

Recall, however, that in this condition we manipulated the expressions with different co-speech gestures to investigate whether the information in gestures (directional or non-directional) could lead listeners to interpret the constructions in a certain way. We now proceed, therefore, to look at the combined effect of verb and gesture type. Figure 3 displays how the interaction of the two variables affects the mean frequency of occurrence of boundary-crossing interpretations. We find that constructions paired with DIRECTIONAL gestures in general receive boundary-crossing interpretations more often than constructions with NON DIRECTIONAL gestures. We also see that the increase, compared to how often the same sentence is interpreted as boundary-crossing when a NON-DIRECTIONAL gesture is present, is particularly large for TYPE 2 verbs.

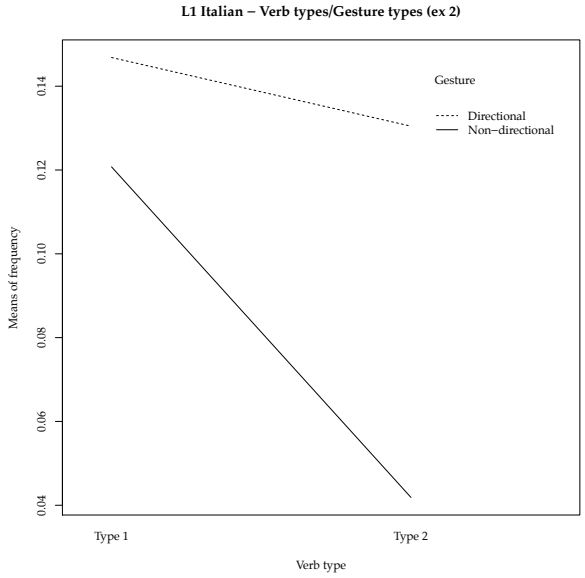
In a repeated measure ANOVA with Verb Type and Gesture Type as within-group variables and Subject as error term, we find main effects for Verb Type and Gesture Type, and an interaction between the two variables. The ANOVA results are reported in table 5. Bonferroni post-tests show pairwise significant

Table 5: Repeated measures 2 x 2 ANOVA results for Verb Type and Gesture Type

Within subject effect	$F(1, 102)$	p	η^2
Verb Type	74.17	< 0.001	0.14
Gesture Type	98.70	< 0.001	0.19
Verb Type * Gesture Type	40.02	< 0.001	0.066

differences between TYPE 1 and TYPE 2 verbs, and between DIRECTIONAL and NON-DIRECTIONAL gestures. To sum up, the co-speech gesture used 'pushes' the interpretation of the motion expression towards the meaning of the gesture itself. When TYPE 2 pure manner verbs are expressed with DIRECTIONAL gestures, the interpretation of these construction are thus much more likely to receive a boundary-crossing interpretation than they would otherwise.

Figure 3: Verb Type and Gesture Type mean freq. of boundary-crossing interpretation



4.3 Comparing NO GESTURE CONDITION with GESTURE CONDITION

Finally, we compare the response patterns from the NO GESTURE CONDITION and the GESTURE CONDITION to investigate whether gestural information potentially can change the interpretation of events compared to the baseline, i.e. whether gestures can maintain or change the interpretation of events. The question is: are there effects of DIRECTIONAL or NON-DIRECTIONAL gestures on the interpretation of TYPE 1 and TYPE 2 verbs across the two conditions? First of all, to rule out a form-based bias caused by two different questionnaire formats (written vs. video), we tested the interpretation of linguistic fillers across the two experiments. In both conditions the filler were ambiguous goal-of-motion constructions, e.g. *Andrea corse a Palermo* - 'Andrea ran to/in Palermo'. In the GESTURE CONDITION these constructions were produced without gestures. We find no difference between formats $\chi^2(1, N = 212) = 0.716$, $p = 0.4$, $\varphi = 0.031$.

The two plots in figure 4 show how verb type and gesture vs no-gesture condition together affect the mean response frequencies of boundary-crossing interpretations. We show this separately for DIRECTIONAL gestures (left-hand side), and NON DIRECTIONAL gestures (right-hand side). The data shows how information displayed through gesture affects the perception of ambiguous motion event expressions. TYPE 1 verbs, which as we saw inherently encode a result feature in their lexical specification, receive the same interpretation in both GESTURE condition involving a DIRECTIONAL gesture and NO GESTURE

condition. On the contrary, when DIRECTIONAL gestures are produced with TYPE 2 verbs, which do not encode a result feature, the interpretation shifts from non-boundary-crossing in the NO GESTURE CONDITION) to a boundary-crossing interpretation. Turning now to the NON DIRECTIONAL gestures, they maintain a non-boundary crossing reading when produced with TYPE 2 verbs, whilst they cause the frequency of boundary-crossing interpretations to decrease when co-occurring with TYPE 1 verbs.

Figure 4: Effect of DIRECTIONAL and NON DIRECTIONAL gestures

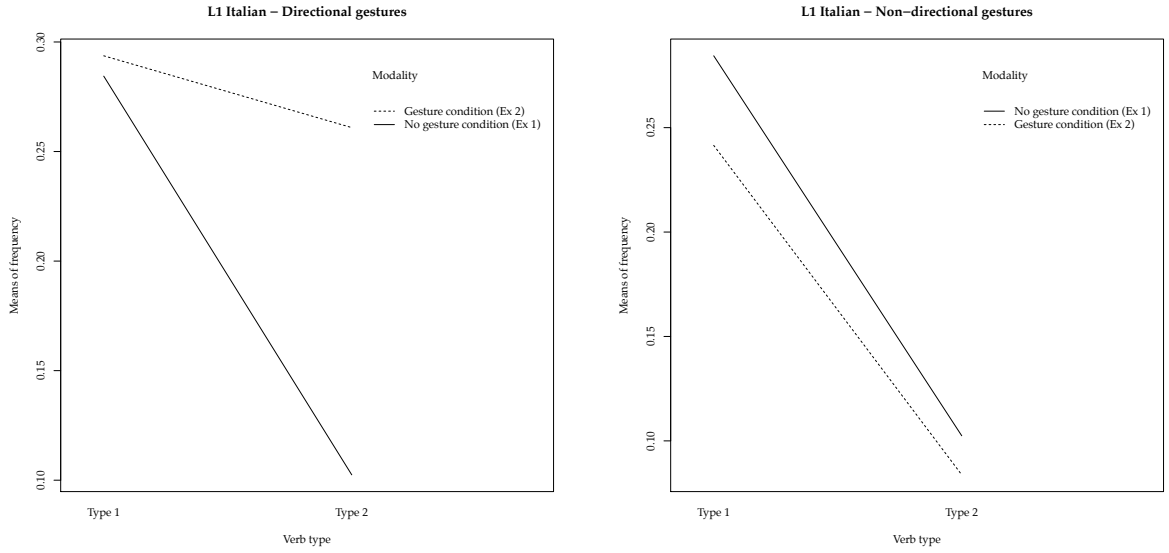


Table 6: Mixed effects 2 x 2 ANOVA results for Experimental Condition and Verb Type (DIRECTIONAL gestures used in the gesture condition)

Main effect	$F(1, 210)$	p	η^2
Experimental Condition (between)	50.59	< 0.001	0.13
Verb Type (within)	143.39	< 0.001	0.199
Experimental Condition * Verb Type	67.85	< 0.001	0.1052

Table 7: Mixed effects 2 x 2 ANOVA results for Experimental Condition and Verb Type (NON-DIRECTIONAL gestures used in the gesture condition)

Main effect	$F(1, 210)$	p	η^2
Experimental Condition (between)	6.99	< 0.01	0.02
Verb Type (within)	357.01	< 0.001	0.39

Mixed effects ANOVAs with Experimental Condition as between-group variable, Verb type as within-group variable and Subject as error term were run to test the significance of these interactions. For the DIRECTIONAL gestures, we found significant main effects for Condition and Verb Type, and a significant interaction between the two. Results are displayed in table 6. Bonferroni post-tests show no significant pairwise difference for TYPE 1 verbs ($p = 0.5$), but it does for TYPE 2 verbs ($p < 0.001$). In constructions with NON-DIRECTIONAL gestures, we found a significant (although smaller) main effect of Condition, a main effect of Verb Type, but no significant interaction between the two $F(1, 210) = 1.83$, $p = 0.18$, $\eta^2 = 0.003$. Results are displayed in table 7. Bonferroni post-tests show a significant pairwise difference in TYPE 1 verbs ($p < 0.01$), but not between TYPE 2 verbs ($p = 0.16$).

5 Discussion

The data from our experimental judgment task provides evidence for the Follian (2008) proposal that Italian manner verbs *can* receive boundary-crossing interpretation when paired with complex locative PPs. From the NO GESTURE CONDITION we found that both with TYPE 1 and TYPE 2 can be interpreted as motion across a spatial boundary, i.e. the figure changes location from one state to another (Berman and Slobin, 1994). However, verbs with an inherent result feature (TYPE 1) are more likely to receive boundary-crossing interpretations than pure manner verbs + complex PP, where, at least according to Bandecchi (2012) neither verb nor any element in the complex PP encode any directional features (see also Iacobini (2014)). In 20% of the cases, however, also examples involving TYPE 2 verbs are given a boundary-crossing interpretation, suggesting that the complex functional structure of the complex PP can give rise to directional interpretation across a boundary even with such verbs, see also Cardini (2012). In essence this finding indicates that the boundary-crossing constraint proposed by Slobin & Hoiting (1994) can be overcome in Italian by using complex PPs to denote a process (path) as well as an end-goal of motion (place). However, since both types of verb + complex PP construction are ambiguous between directional and locative meaning, many participants also interpreted them as purely locative motion (e.g. as motion within a container). As stated in the introduction, the manner verb + complex PP construction is not widely used in Italian to express boundary-crossing movement, yet modern spoken Italian has seen an increased tendency to express directional motion in non-boundary-crossing situations with manner verbs and path-denoting satellites (e.g. *rotola giù* - 'rolls down') (Hijazo-Gascón and Ibarretxe-Antuñano, 2013; Iacobini and Masini, 2006; Wessel-Tolvig and Paggio, 2016). Deeply rooted thinking-for-speaking patterns (Slobin, 1996), linguistic habits (Cardini, 2008) and a tendency to avoid ambiguity may all contribute to Italian speakers not choosing these atypical, yet grammatically available manner verb + complex PP constructions.

In the GESTURE CONDITION we tested the effects of gestural information when interpreting the manner verb + complex constructions. Listeners integrate the information to build a more complete picture of the expressed situation (Kelly et al., 2010; Kelly et al., 2014) especially in those situations where the content is ambiguous and there are no anaphoric cues to a context (Holle and Gunter, 2007; Dick et al., 2009). Listeners are influenced by the additional information provided by gestures and the interpretation of the expressions is shifted towards their content. When the speaker in the videos used a DIRECTIONAL gesture with the manner verb + complex construction, listeners were more likely to interpret the expression as movement across a boundary as opposed to movement that did not cross any spatial markers. This effect was stronger for TYPE 2 verbs, where it can be said that the default locative interpretation of the verb is overridden. NON DIRECTIONAL gestures, similarly, increased the probability of a non-boundary crossing interpretation for both verb types, however the effect was stronger with TYPE 1 verbs. In essence, the most striking effect of gestures occurs when there is some sort of semantic 'incongruence' between speech and gesture information.

When we compared the response patterns across conditions we found that the strongest effects are found when we combine DIRECTIONAL gestures with TYPE 2 verbs ($p < 0.001$) and NON DIRECTIONAL gestures with TYPE 1 verbs ($p < 0.01$). DIRECTIONAL gestures significantly increased boundary-crossing interpretations and NON DIRECTIONAL gestures significantly decreased boundary-crossing interpretation. The other two combinations maintained scores of interpretation across conditions (*ns).

To summarize our findings, the study confirms earlier claims that boundary-crossing interpretations of the manner verb + complex PP construction are possible in Italian (Folli and Ramchand, 2005; Cardini, 2012), but more importantly, we find that the information provided by gestures can affect the interpretation of these ambiguous expressions and 'override' the default meaning expressed only in speech. The data thus provides important knowledge of how listeners integrate gestural information in comprehension to derive speaker meanings and thus construct a more complete picture of the content of an utterance (McNeill, 1992; McNeill, 2000).

Acknowledgements

We would like to thank Maria Grazia Busà and Alice Cravotta at LCL in Padova (Italy) for help with data collection, Lorenzo Menon at University of Copenhagen for his acting in the elicitation materials, and all the subjects who participated in the studies. This research was funded by the Danish Council for Independent Research.

References

Jon Aske. 1989. Path predicates in English and Spanish: A closer look. In *Berkeley Linguistic Society 15*, Berkeley, USA.

Panos Athanasopoulos and Emanuel Bylund. 2013. The 'thinking' in thinking-for-speaking: Where is it? *Language, Interaction & Acquisition*, 4(1):91–100.

Valeria Bandecchi. 2012. Prepositional phrases and manner-of-motion verbs in Italian. In *22nd Colloquium on Generative Grammar*, Barcelona, Spain.

Geoffrey Beattie and Heather Shovelton. 1999. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? an experimental investigation. *Semiotica*, 123(1-2):1–30.

John Beavers, Beth Levin, and Shiao Wei Tham. 2010. The typology of motion expressions revisited. *Journal of Linguistics*, 46(2):331–377.

Ruth Aronson Berman and Dan Isaac Slobin. 1994. *Relating events in narrative: A crosslinguistic developmental study*. Psychology Press, Hillsdale, NJ.

Teresa Cadierno and Lucas Ruiz. 2006. Motion events in Spanish l2 acquisition. *Annual Review of Cognitive Linguistics*, 4:183–216.

Teresa Cadierno, 2012. *Thinking for speaking in second language acquisition*. Wiley-Blackwell.

Filippo-Enrico Cardini. 2008. Manner of motion saliency: An inquiry into Italian. *Cognitive Linguistics*, 19(4):533–569.

Filippo-Enrico Cardini. 2012. Grammatical constraints and verb-framed languages: The case of Italian. *Language and Cognition*, 4(3):167–201.

William Croft, Jhanna Bardal, Willem Hollmann, Violeta Sotirova, and Chiaki Taoka, 2010. *Revising Talmys typological classification of complex event constructions*, pages 201–235. John Benjamins.

Anthony Steven Dick, Susan Goldin-Meadow, Uri Hasson, Jeremy I. Skipper, and Steven L. Small. 2009. Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human brain mapping*, 30(11):3509–3526.

Raffaella Folli and Gillian Ramchand, 2005. *Prepositions and Results in Italian and English: An Analysis from Event Decomposition*, volume 32 of *Studies in Theoretical Psycholinguistics*, book section 5, pages 81–105. Springer Netherlands.

Raffaella Folli, 2008. *Complex PPs in Italian*, pages 197–221. John Benjamins Publishing Company.

Berit Gehrke. 2007. On directional readings of locative prepositions. In *ConSOLE XIV*, pages 99–120, Vitoria-Gasteiz.

Whitney Goodrich Smith and Carla Hudson Kam. 2012. Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun comprehension. *Language & Cognition (De Gruyter)*, 4(2):75–98.

Whitney Goodrich Smith and Carla Hudson Kam. 2015. Children's use of gesture in ambiguous pronoun interpretation. *Journal of Child Language*, 42(03):591–617.

Marianne Gullberg, 2011. *Thinking, speaking and gesturing about motion in more than one language*, book section 5, pages 143–169. Multilingual Matters.

Simon Harrison. 2011. The creation and implementation of a gesture code for factory communication. In *GESPIN 2011 Gesture Conference*, Bielefeld, Germany.

Alberto Hijazo-Gascón and Iraide Ibarretxe-Antuñano, 2013. *Same family, different paths*. John Benjamins.

Henning Holle and Thomas C. Gunter. 2007. The role of iconic gestures in speech disambiguation: Erp evidence. *Journal of Cognitive Neuroscience*, 19(7):1175–1192.

Claudio Iacobini and Francesca Masini. 2006. The emergence of verb-particle constructions in Italian: locative and actional meanings. *Morphology*, 16:155–188.

Claudio Iacobini and Carla Vergaro. 2014. The role of inference in motion event encoding / decoding: a cross-linguistic inquiry into English and Italian. *Lingue e linguaggio*, 2:211–240.

Spencer Kelly, Asli Özyürek, and Eric Maris. 2010. Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2):260–267.

Spencer Kelly, Meghan L Healey, Asli Özyürek, and Judith Holler. 2014. The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, 22(2).

Adam Kendon, 1980. *Gesture and speech: two aspects of the process of utterance*, pages 207–227. Mouton.

Sotaro Kita and Asli Özyürek. 2003. 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*, 48(1):16–32.

Sotaro Kita, Asli Özyürek, Shanley Allen, Amanda Brown, Reyhan Furman, and Tomoko Ishizuka. 2007. Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, 22(8):1212–1236.

Jaume Mateu and Gemma Rigau. 2010. Verb-particle constructions in romance: A lexical-syntactic account. *Probus*, 22(2):241–269.

David McNeill. 1992. *Hand and mind: what gestures reveal about thought*. University of Chicago Press, Chicago.

David McNeill. 2000. *Language and gesture*. Cambridge University Press, Cambridge, UK.

David McNeill. 2005. *Gesture & Thought*. The University of Chicago Press, Chicago.

Asli Özyürek, Sotaro Kita, Shanley Allen, Reyhan Furman, and Amanda Brown. 2005. How does linguistic framing of events influence co-speech gestures?: Insights from crosslinguistic variations and similarities. *Gesture*, 5(1-1):219–240.

Dan Isaac Slobin and N Hoiting. 1994. Reference to movement in spoken and signed languages: Typological considerations. In *Twentieth Annual Meeting of the Berkeley Linguistics Society*, pages 487–505, Berkeley, USA.

Dan Isaac Slobin. 1987. Thinking for speaking. In *Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pages 435–444, Berkeley, USA.

Dan Isaac Slobin, 1996. *From thought and language to thinking for speaking*, pages 70–96. Cambridge University Press.

Dan Isaac Slobin, 2004. *The many ways to search for a frog: Linguistic typology and the expression of motion events*, pages 219–257. Lawrence Erlbaum Associates.

Gale Stam. 2006. Thinking for speaking about motion: L1 and l2 speech and gesture. *International Review of Applied Linguistics*, 44:143–169.

Leonard Talmy, 1985. *Semantics and syntax of motion*, volume 3, pages 57–149. Cambridge UniversityPress.

Leonard Talmy. 1991. Path to realization: A typology of event conflation. In *Seventeenth Annual Meeting of the Berkeley Linguistics Society*, pages 480–519, Berkeley, USA.

Leonard Talmy. 2000. *Toward a Cognitive Semantics: Typology and Process in Concept Structuring*, volume II. The MIT Press, Cambridge.

Bjørn Wessel-Tolvig and Patrizia Paggio. 2016. Revisiting the thinking-for-speaking hypothesis: Speech and gesture representation of motion in Danish and Italian. *Journal of Pragmatics*, 99:39–61.

Bjørn Wessel-Tolvig. 2015. Breaking boundaries: How gestures reveal conceptualization of boundary-crossing in Italian. In *Gespin 4*, Nantes, France.

BCI effectiveness test through N400 replication study

Thomas Ousterhout
University of Copenhagen / Njalsgade 136, building 27, 2300 Copenhagen S
tko@hum.ku.dk

Abstract

In an effort to test the ability of a commercial grade EEG headset to effectively measure the N400 ERP, a replication study was conducted to see if similar results could be produced as that which used a medical grade EEG. Pictures of meaningful and meaningless hand postures were borrowed from the author of the replicated study and subjects were required to perform a semantic discrimination task. The N400 was detected indicating semantic processing of the meaningfulness of the hand postures. The results corroborate those of the replication study and support the use of some commercial grade EEG headsets for non-critical research applications.

1 Introduction

This study was designed to promote and validate the functionality of commercially available and user friendly neuroimaging technology as a brain-computer interface (BCI) and Electroencephalography (EEG) research tool. Developments in cognitive technologies are allowing researchers and users to access cognitive information in a cost effective manner. EEG is a measurement tool used to detect and measure the electrical signals in the brain when neurons communicate with each other. While invasive, cortically-implanted electrodes, allow for a more precise method of measuring brain activity, non-invasive scalp electrodes allow for a much more appropriate scientific method for the average researcher and user (Lin et al., 2008).

BCIs have shown an incredible ability to allow those with mobility disabilities to control medical devices such as prosthetic limbs (Nunez and Srinivasan, 2006; Guger et al., 1999; Müller-Putz and Pfurtscheller, 2008; Farwell and Donchin, 1988), wheelchairs (Barea et al., 2002a; Barea et al., 2002b; Rebsamen et al., 2006; Rebsamen et al., 2007; Barea et al., 2003; Chowdhury and Shakim, 2014) and robots (Neto et al., 2006; Chowdhury et al., 2014; Tripathy and Raheja, 2015). One reason this is possible is due to the ability to predict voluntary human movement more than a second before it occurs (Bai et al., 2011; Funase et al., 1999; Morash et al., 2008).

However, BCIs are not just used for mind controlled vehicles or devices using cognitive thought, they can also perform as diagnostic tools to detect driver fatigue (Zhao et al., 2011; Lin et al., 2008; Jap et al., 2009) and drowsiness (De Rosario et al., 2010; Khushaba et al., 2011; Lin et al., 2010; Eoh et al., 2005). This shows that applications using BCIs range from medical purposes for people who are locked in a vegetative state and helping them communicate with the world, to gaming/recreational purposes for healthy users who want to enhance their lives with smart technology.

While many EEG and BCI systems use medical grade technology as a data acquisition tool, the relatively cheap and wireless BCI system called the Emotiv EPOC is a cost effective consumer grade EEG unit with only 14 channels and this system has proven effective in several studies (Ousterhout and Dyrholm, 2013; Debener et al., 2012; De Vos et al., 2014a; Campbell et al., 2010; De Vos et al., 2014b). However the technology is still controversial as there are some studies that do not support its use fully (Duvinage et al., 2012; Stytsenko et al., 2011; Liu et al., 2012; Duvinage et al., 2013), stating that the

system, being significantly worse than standard medical grade EEG, should only be used in noncritical applications. One noncritical application could certainly be communication.

When people communicate face-to-face, they typically engage in multimodal communication which simultaneously uses both modes of auditory and visual information. Auditory information normally only consists of speech, and visual information can include things like body behavior such as facial expressions, hand and arm gestures, and body posture. Visual information is also used inter alia to disambiguate context by providing supplemental information to the dominantly used vocal information, can be used instead of speech, and can change the meaning of speech (Goldin-Meadow, 1999; Kelly et al., 1999; McNeill, 2008; Kendon, 2004). While the auditory modality provides the most information content in face-to-face communication, thus typically being the dominant modality of communication, the visual cues are very important and sometimes necessary to understand fully what the intended message is (Clark, 1996).

Hand gesturing, for example, is an integral part of our daily communication paradigm. Hand gesture types can be categorized into several groups, while simultaneously being part of a larger continuum. For example, one type is called an emblem, which is a hand gesture requiring no verbal supplement, and has a conventionalized meaning in a particular culture, such as the thumbs up gesture in western cultures. These gestures can be useful in face-to-face communication because one gesture alone can give a complicated message to the recipient instantaneously (McNeill, 1992). Therefore emblems can be considered unspoken words or phrases, since there is a strong relationship between the gesture and it's meaning. Another type are iconic gestures, which are used to symbolize something, such as putting one's hands in the shape of a ball when talking about a ball. There are also deictic gestures, which comprise pointing hand postures.

Gestures thus contribute to the semantics of the dialogue in face-to-face communication. Semantics can be measured with EEG by looking at event-related potentials (ERPs), which are amplitude deflections in the brain produced in response to certain events or stimuli. One ERP has been studied for over thirty years to measure semantic processing (Kutas and Federmeier, 2011; Duncan et al., 2009; Kutas and Federmeier, 2000; Gunter and Bach, 2004) called the N400 ERP which is a negative deflecting component occurring 400 ms after the onset of a auditory or visual stimulus.

This ERP is used to measure semantic congruence. More specifically, the presence of an incongruous stimulus results in a much larger negative deflection than that of a congruous one, or one that is expected. Therefore, we can measure whether texts, images, or any other stimulus type are congruous or incongruous with the preceding context by looking at this ERP amplitude deflection. However, not all electrode positions measure this ERP since the responses to abstract words in semantic processing are typically found in centro-parietal sites, while concrete words, such as ones referring to picturable objects, have a frontal distribution (Holcomb et al., 1999).

The N400 has also shown utility in its ability to measure the amount of cognitive load required for an individual in semantic memory retrieval. This is because the ability to process the information from probe stimuli is highly dependent on one's ability to recall previous relevant stimuli from any of the multimodal channels such as images or sounds. This difficulty, or cognitive load, is associated with memory representations and cues from previous content priming the meaningful probe stimulus (Federmeier and Kutas, 2001; Lau et al., 2008; Van Petten and Luka, 2006). Therefore, when a difficult stimulus requires more effort to process, thus having more cognitive load, the N400's amplitude deflection is larger than when it is easy. It is therefore that the N400 is larger for rarely used words and when semantically incongruent or unrelated to previously acquired content (PETTEN, 1995; Laszlo and Federmeier, 2011).

A study done by Gunter and Bach (2004), investigated the N400 effect using pictures of semantically meaningful and meaningless hand postures. Pictures of 11 common and well-known emblematic, iconic, and deictic gestures were used as the meaningful stimuli along with 11 similarly positioned yet meaningless hand positions. The meaningful hand postures include things like "thumbs up" and "peace/victory" and other familiar postures which have a symbolic meaning to a given culture. During the pictorial semantic categorization task, subjects were required to identify, through a button press response, if each randomly displayed picture was meaningful or meaningless. They found that in comparison to mean-

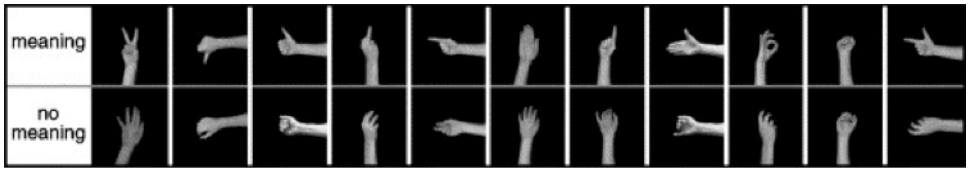


Figure 1: The 11 meaningful and similarly positioned yet meaningless hand postures provided by Gunter and Bach (2004).

ingful hand positions, the meaningless ones produced a larger negative going amplitude deflection in the centro-parietal region, which they classified as the N400.

This current study presumes to replicate precisely the study done by Gunter and Bach (2004) in an attempt to also find the N400, despite the fact that the Emotiv has no electrodes positioned that can measure the centro-parietal region which is where the N400 originates from. The hypothesis is that since the N400 is such a large ERP, even with the poor resolution of the 14-channel Emotiv, in comparison to the high resolution 59-channel medical grade EEG scalp cap used by Gunter and Bach (2004), the Emotiv will still be able to detect the N400 from the meaningless hand postures.

2 Method

2.1 Participants

This study used 16 participants who were native English or fluent English speaking adults at the University of Copenhagen. Their ages ranged from 20-37 years (mean = 26.9), 9 were males and all were right handed. All participants signed an informed consent form ensuring their understanding of the experiment to be conducted. All participants had normal or correct-to-normal vision with no reported psychiatric, neurological, or reading disorders that could disrupts this study's efficacy.

2.2 Stimuli

Participants were presented with stimuli courtesy of Gunter and Bach (2004) which consisted of 66 meaningful and 66 meaningless grey-scale hand posture photos. Each of the 11 meaningful and meaningless hand postures seen in Figure 1 were photographed by six different people and all 132 pictures were shown in 3 cycles.

2.3 Procedure

Using Paradigm stimulation software, which is a stimulus presentation software program that is good a millisecond display timing, a trial of the discrimination task progressed first with a random hand posture for 700 ms, then a blank screen for 500 ms, and finally a GO signal was presented indicating that the participant had to input with a button press if they judged the hand posture as meaningful or meaningless. This lasted approximately 20-25 minutes for each participant. Since the Emotiv is designed for real world applications, the study was done in a closed university office with normal lighting conditions and the possibility for auditory noise outside.

2.4 Electrophysiological Acquisition

For EEG acquisition, the 14-channel Emotiv was used which has electrodes at the International 10/20 system at AF3, F7, F3, FC5, T7, P7, O1,O2, P8, T8, FC6, F4, F8, AF4 with two left and right mastoid references at P3 and P4. The data was filtered offline from 0.1 to 30 HZ and sampled continuously at 128 HZ. To support the use of the Emotiv in the real world involving noisy environments in real time, no artifact rejection or correction was applied, however only correct responses were used. ERPs were identified and measured off-line using Matlab's ERPLab with a baseline averaged from the -200 to stimulus onset interval window and average ERPs lasted 1000 ms after the onset of the probe.

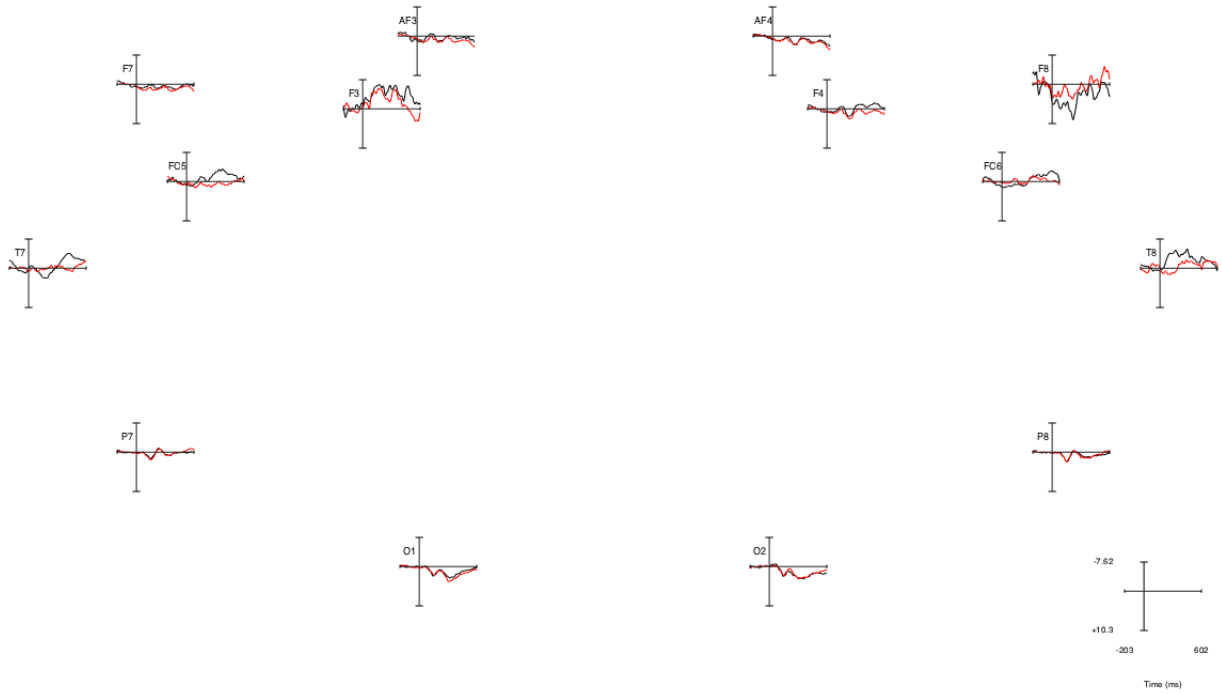


Figure 2: Wavelengths from the time window of -203 to 602 ms, where negativity is plotted upwards, and with congruous (red) and incongruous (black) conditions.

3 Results

3.1 Electrophysiological Results

The ERPs were measured using a repeated measures ANOVA with a 2 x 6 design (meaningfulness x electrode) using only the 6 electrodes F3, FC5, T7, T8, FC6 and F4 since they were closest to the PZ electrode position which is typically used to measure the N400. The mean amplitude for these 6 electrodes was calculated within the time window of 300-500 ms after stimulus onset. Figure 2 shows the grand average wavelengths of meaningful and meaningless stimuli and Figure 3 shows the scalp map distribution in the measurement time window in 50 ms intervals. The ANOVA Sphericity Assumed test showed an effect for meaningfulness ($F(1,15) = 5.36$, $p = 0.035$) and thus was identified as an N400.

4 Discussion

In summary, this study investigated the N400 effect regarding meaningless hand gestures compared to meaningful hand gestures made up of emblem, iconic, and deictic gestures. This study also replicated the paradigm and reproduced the results of Gunter and Bach (2004) regarding N400 detection. Most importantly, this study gives further evidence to support the use of a simple and affordable BCI as a research and user tool for noncritical EEG/ERP/BCI applications.

This study further corroborates with previous research regarding the issue of if some meaningful hand postures, such as emblems, are lexicalized and thus processed in the brain like words are. The theory was that since the comparison between meaningful words and similar yet false pseudo words produces an N400 effect, the same would go for meaningful hand gestures and similar yet false meaningless hand gestures. The increased N400 of the meaningless hand postures in comparison to the meaningful ones is similar to other studies (Bentin, 1986; Bentin et al., 1985).

The only difference between the result of this study and those done by Gunter and Bach (2004) are that there was an N300 effect with right-frontal distribution in that study which is indicative of picture processing and thus should have also been seen in this study (Barrett and Rugg, 1990; Federmeier and Kutas, 2001; McPherson and Holcomb, 1999; West and Holcomb, 2002). This current study found a greater negativity lateralized towards the left. The cause of this difference is unknown but will be investigated.

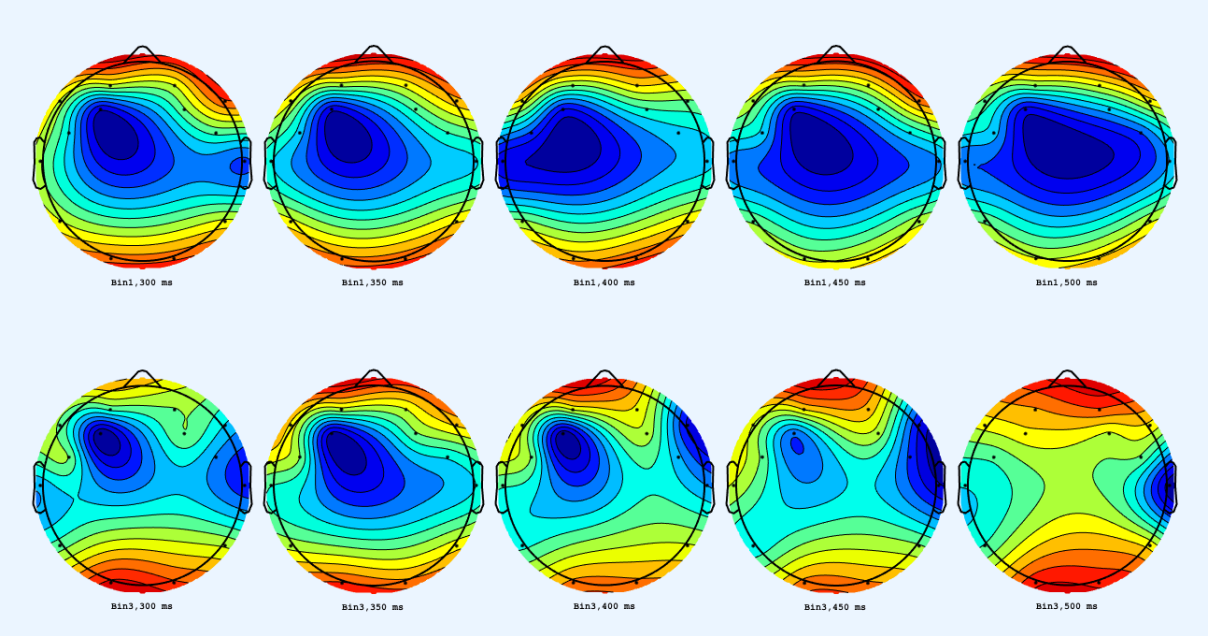


Figure 3: Scalp maps of incongruous (top) and congruous (bottom) conditions from 300 to 500 ms in 50 ms intervals.

However, Gunter and Bach (2004) did mention that the N300 and N400 effects were relatively small and could have been due to the large repetition of stimuli in the experiment and also could potentially have been facilitated with priming. Another potential explanation for the difference could be that Gunter and Bach (2004) used 22 native-German speaking students, where this study used 16 students from countries all over the world. This cultural difference could have had a dramatic effect on the semantic processing of the meaningful hand postures.

However, the most important part of this study is the demonstration that a simple, cost-effective, 14-channel EEG headset can detect the N400 in a similar manner that medical grade 59-channel EEG systems can. Even more impressive is that the system worked without any electrodes covering the source of the N400, and the data acquisition was done in a regular room with real world auditory and visual distractions, and no type of artifact rejection or correction was done. This further supports the usefulness of the commercially available EEG equipment, such as the Emotiv, as a research tool for ERP detection and user interface for BCIs.

5 Conclusions

This article demonstrated through a replication study of Gunter and Bach (2004) that the Emotiv headset, which is a relatively cheap commercial grade EEG acquisition device, can give results comparable to those of expensive high resolution medical grade equipment when measuring some ERP signals such as the N400. This is useful for a number of reasons. First, researchers and students can perform EEG experiments and contribute to the scientific world through published scholarly articles without the need of a medical environment and expensive equipment.

This also means that the rather exclusive field of EEG research can be easily accessed by virtually anyone allowing the volume of contribution to the field to increase dramatically. Furthermore, it can be assumed that the participants in the EEG experiments behave more naturally in an office than in an experimental laboratory and when they wear a simple EEG headset. Further research could be seeing what other areas the Emotiv is comparable to expensive EEG equipment or using the Emotiv for more semantic priming and N400 experiments.

6 Acknowledgements

I would like to thank Thomas Gunther for the permission to use his stimuli in my study.

References

Ou Bai, Varun Rathi, Peter Lin, Dandan Huang, Harsha Battapady, Ding-Yu Fei, Logan Schneider, Elise Houdayer, Xuedong Chen, and Mark Hallett. 2011. Prediction of human voluntary movement before it occurs. *Clinical Neurophysiology*, 122(2):364–372.

Rafael Barea, Luciano Boquete, Manuel Mazo, and E López. 2002a. Wheelchair guidance strategies using eog. *Journal of intelligent and robotic systems*, 34(3):279–299.

Rafael Barea, Luciano Boquete, Manuel Mazo, and Elena López. 2002b. System for assisted mobility using eye movements based on electrooculography. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 10(4):209–218.

Rafael Barea, Luciano Boquete, Luis Miguel Bergasa, Elena López, and Manuel Mazo. 2003. Electro-oculographic guidance of a wheelchair using eye movements codification. *The International Journal of Robotics Research*, 22(7-8):641–652.

Sarah E Barrett and Michael D Rugg. 1990. Event-related potentials and the semantic matching of pictures. *Brain and cognition*, 14(2):201–212.

Shlomo Bentin, Gregory McCarthy, and Charles C Wood. 1985. Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, 60(4):343–355.

S Bentin. 1986. Visual word perception and semantic processing: an electrophysiological perspective. *Israel journal of medical sciences*, 23(1-2):138–144.

Andrew Campbell, Tanzeem Choudhury, Shaohan Hu, Hong Lu, Matthew K Mukerjee, Mashfiqui Rabbi, and Rajeev DS Raizada. 2010. Neurophone: brain-mobile phone interface using a wireless eeg headset. In *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds*, pages 3–8. ACM.

Pritom Chowdhury and SS Kibria Shakim. 2014. *Optimizing cognitive efficiency of emotiv EPOC and controlling wheelchair through it*. Ph.D. thesis, BRAC University.

Pulak Chowdhury, SS Kibria Shakim, Muhammad Rezaul Karim, and Md Khalilur Rhaman. 2014. Cognitive efficiency in robot control by emotiv epoc. In *Informatics, Electronics & Vision (ICIEV), 2014 International Conference on*, pages 1–6. IEEE.

Herbert H Clark. 1996. *Using language*, volume 1996. Cambridge university press Cambridge.

Helios De Rosario, Jose S Solaz, N Rodrıguez, and Luis M Bergasa. 2010. Controlled inducement and measurement of drowsiness in a driving simulator. *IET intelligent transport systems*, 4(4):280–288.

Maarten De Vos, Katharina Gandras, and Stefan Debener. 2014a. Towards a truly mobile auditory brain–computer interface: Exploring the p300 to take away. *International Journal of Psychophysiology*, 91(1):46–53.

Maarten De Vos, Markus Kroesen, Reiner Emkes, and Stefan Debener. 2014b. P300 speller bci with a mobile eeg system: comparison to a traditional amplifier. *Journal of neural engineering*, 11(3):036008.

Stefan Debener, Falk Minow, Reiner Emkes, Katharina Gandras, and Maarten Vos. 2012. How about taking a low-cost, small, and wireless eeg for a walk? *Psychophysiology*, 49(11):1617–1621.

Connie C Duncan, Robert J Barry, John F Connolly, Catherine Fischer, Patricia T Michie, Risto Näätänen, John Polich, Ivar Reinvang, and Cyma Van Petten. 2009. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, p300, and n400. *Clinical Neurophysiology*, 120(11):1883–1908.

Matthieu Duvinage, Thierry Castermans, Thierry Dutoit, M Petieau, T Hoellinger, C De Saedeleer, K Seetharaman, and G Cheron. 2012. A p300-based quantitative comparison between the emotiv epoc headset and a medical eeg device. *Biomedical Engineering*, 765:2012–764.

Matthieu Duvinage, Thierry Castermans, Mathieu Petieau, Thomas Hoellinger, Guy Cheron, and Thierry Dutoit. 2013. Performance of the emotiv epoc headset for p300-based applications. *Biomed Eng Online*, 12:56.

Hong J Eoh, Min K Chung, and Seong-Han Kim. 2005. Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. *International Journal of Industrial Ergonomics*, 35(4):307–320.

Lawrence Ashley Farwell and Emanuel Donchin. 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523.

Kara D Federmeier and Marta Kutas. 2001. Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1):202.

Arao Funase, Tohru Yagi, Yoslhiaki Kuno, and Yoshiki Uchikawa. 1999. Prediction of eye movements from eeg. In *Neural Information Processing, 1999. Proceedings. ICONIP'99. 6th International Conference on*, volume 3, pages 1127–1131. IEEE.

Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429.

Christoph Guger, Werner Harkam, Carin Hertnaes, and Gert Pfurtscheller. 1999. Prosthetic control by an eeg-based brain-computer interface (bci). In *Proc. aaate 5th european conference for the advancement of assistive technology*, pages 3–6. Citeseer.

Thomas C Gunter and Patric Bach. 2004. Communicating hands: Erps elicited by meaningful symbolic hand postures. *Neuroscience Letters*, 372(1):52–56.

Phillip J Holcomb, John Kounios, Jane E Anderson, and W Caroline West. 1999. Dual-coding, context-availability, and concreteness effects in sentence comprehension: an electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(3):721.

Budi Thomas Jap, Sara Lal, Peter Fischer, and Evangelos Bekiaris. 2009. Using eeg spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications*, 36(2):2352–2359.

Spencer D Kelly, Dale J Barr, R Breckinridge Church, and Katheryn Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4):577–592.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Rami N Khushaba, Sarath Kodagoda, Sara Lal, and Gamini Dissanayake. 2011. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *Biomedical Engineering, IEEE Transactions on*, 58(1):121–131.

Marta Kutas and Kara D Federmeier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences*, 4(12):463–470.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp). *Annual review of psychology*, 62:621.

Sarah Laszlo and Kara D Federmeier. 2011. The n400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, 48(2):176–186.

Ellen F Lau, Colin Phillips, and David Poeppel. 2008. A cortical network for semantics:(de) constructing the n400. *Nature Reviews Neuroscience*, 9(12):920–933.

By Chin-Teng Lin, Li-Wei Ko, Jin-Chern Chiou, Jeng-Ren Duann, Ruey-Song Huang, Sheng-Fu Liang, Tzai-Wen Chiu, and Tzyy-Ping Jung. 2008. Noninvasive neural prostheses using mobile and wireless eeg. *Proceedings of the IEEE*, 96(7):1167–1183.

Chin-Teng Lin, Che-Jui Chang, Bor-Shyh Lin, Shao-Hang Hung, Chih-Feng Chao, and I-Jan Wang. 2010. A real-time wireless brain–computer interface system for drowsiness detection. *Biomedical Circuits and Systems, IEEE Transactions on*, 4(4):214–222.

Yue Liu, Xiao Jiang, Teng Cao, Feng Wan, Peng Un Mak, Pui-In Mak, and Mang I Vai. 2012. Implementation of ssvep based bci with emotiv epoc. In *Virtual Environments Human-Computer Interfaces and Measurement Systems (VECIMS), 2012 IEEE International Conference on*, pages 34–37. IEEE.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

David McNeill. 2008. *Gesture and thought*. University of Chicago Press.

W Brian McPherson and Phillip J Holcomb. 1999. An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, 36(01):53–65.

Valerie Morash, Ou Bai, Stephen Furlani, Peter Lin, and Mark Hallett. 2008. Classifying eeg signals preceding right hand, left hand, tongue, and right foot movements and motor imageries. *Clinical neurophysiology*, 119(11):2570–2578.

Gernot R Müller-Putz and Gert Pfurtscheller. 2008. Control of an electrical prosthesis with an ssvep-based bci. *Biomedical Engineering, IEEE Transactions on*, 55(1):361–364.

Anselmo Frizera Neto, Wanderley Cardoso Celeste, Vinicius Ruiz Martins, et al. 2006. Human-machine interface based on electro-biological signals for mobile vehicles. In *Industrial Electronics, 2006 IEEE International Symposium on*, volume 4, pages 2954–2959. IEEE.

Paul L Nunez and Ramesh Srinivasan. 2006. *Electric fields of the brain: the neurophysics of EEG*. Oxford university press.

Thomas Ousterhout and Mads Dyrholm. 2013. Cortically coupled computer vision with emotiv headset using distractor variables. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 245–250. IEEE.

CYMA PETTEN. 1995. Words and sentences: Event-related brain potential measures. *Psychophysiology*, 32(6):511–525.

Brice Rebsamen, Etienne Burdet, Cuntai Guan, Haihong Zhang, Chee Leong Teo, Qiang Zeng, Marcelo Ang, and Christian Laugier. 2006. A brain-controlled wheelchair based on p300 and path guidance. In *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, pages 1101–1106. IEEE.

Brice Rebsamen, Chee Leong Teo, Qiang Zeng, Marcelo H Ang Jr, Etienne Burdet, Cuntai Guan, Haihong Zhang, and Christian Laugier. 2007. Controlling a wheelchair indoors using thought. *Intelligent Systems, IEEE*, 22(2):18–24.

Kirill Stytsenko, Evaldas Jablonskis, and Cosima Prahm. 2011. Evaluation of consumer eeg device emotiv epoc. In *MEi: CogSci Conference 2011, Ljubljana*.

Devashree Tripathy and Jagdish Lal Raheja. 2015. Design and implementation of brain computer interface based robot motion control. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, pages 289–296. Springer.

Cyma Van Petten and Barbara J Luka. 2006. Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and language*, 97(3):279–293.

W Caroline West and Phillip J Holcomb. 2002. Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, 13(3):363–375.

Chunlin Zhao, Chongxun Zheng, Min Zhao, Yaling Tu, and Jianping Liu. 2011. Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. *Expert Systems with Applications*, 38(3):1859–1865.

Investigation of the semantic priming effect with the N400 using symbolic pictures in text

Thomas Ousterhout
University of Copenhagen / Njalsgade 136, building 27, 2300 Copenhagen S
tko@hum.ku.dk

Abstract

In face-to-face communication, a large portion of communicative devices rely on the visual modality of bodily behaviors which include facial expression and hand gestures. However through the use of digitally mediated communication which is becoming increasingly prevalent with advances in technology, people are evolving their way to communicate. Texts become shorter and the use of emojis are changing. Facial emojis are symbols for human faces that have become increasingly popular with communicative devices. The original and still most frequent use of emojis is to provide a comment to the text which they follow. However, the latest trend is also to use emojis in the middle of sentences replacing words or adding information to the text. Through the use of EEG and the N400 ERP component, this study investigates which objects emojis refer to via an internet survey and a EEG semantic priming test in which moving emojis in sentences are paired with congruous and incongruous probes. The results of both the survey and the EEG test indicate that there is no preference for particular positions of the emojis and that some of the unusual emojis were ambiguous and did not add to comprehension.

1 Introduction

In order to interact with people, no matter what the social circumstance or environment is, it is absolutely necessary to understand what others are doing, intending and feeling. Without the ability to understand others, intended and unintended messages would be lost in the process of communicating and cooperation would be unproductive. There are certainly many parts of the brain that are responsible for effective communication, one part in particular is called the mirror mechanism (Rizzolatti and Fabbri-Destro, 2008).

Mirror neurons, which were first discovered in monkeys, have also been found in humans. Whenever an individual sees an individual performing a motor action, mirror neurons activate a part of the brain that also fire when the observer executes the exact same action themselves. (Jeannerod, 1994) believes this is for learning purposes. As many are familiar with, students watching a teacher, rather than just listening, help with the learning process of performing the action or task. This is because while watching the agent perform the task, the mirror neurons encode a representation of the action itself, which it just has to repeat when executing that action.

(Rizzolatti et al., 1996) theorize that the mirroring system contributes to understanding motor actions through recognition, differentiation, and knowing how to respond appropriately. Simply put, when an agent performs an action, it can be assumed that there is an intention with a prediction of a specific outcome. An observer can learn quickly how to produce specific outcomes from observation alone rather than practice through this mirror mechanism and more importantly what the meaning of those actions represent.

Interestingly, there is a large amount of generalization when it comes to the type of stimuli in which mirror neurons respond to. (Rizzolatti and Craighero, 2004) found that the same mirror neuron cluster

fires when a human grasps an object as well as when a monkey does the exact same performance. This same firing pattern also happens when watching the action from a distance or at proximity.

While mirror neurons are very important in action learning and understanding, another theory by (MacNeilage, 1998) proposes that human speech evolved from monkey open-close jaw movements such as when they perform lipsmacks. These simple facial manipulations created a type of faciovisual communication, also known currently as facial expressions or gestures. This suggests that communication began as a visual modality which later was supplemented with sounds. These are all proposals of why mirror neurons play a role in matching observed and executed actions and why they are important in understanding each others behavior. Also mirror neurons are found in Brocas area, which is significantly responsible for understanding speech, and furthermore several theories address how speech evolved from visual/gestural communication (Armstrong et al., 1995; Rizzolatti and Arbib, 1998; Corballis, 2002).

Emotion expression and comprehension are also crucial communicative devices needed for effective message transmission. (Singer et al., 2004; Wicker et al., 2003) both performed experiments in which participants experienced an emotion such as disgust or pain, and then watched another go through the same experience. Both studies showed that similar neurological activity was produced in experience and observation conditions suggesting a mirror mechanism involved.

It seems however that observing an action performed by an agent is not the only factor activating the mirror neurons, but rather perceiving the action performed by the agent. Observation is typically required in order to produce some kind of understanding, however (Kohler et al., 2002; Umilta et al., 2001) removed the ability of the participants to produce visual observation and yet still were able to measure the variable of understanding in isolation. They set up an experiment where visual observation was not possible due to a blocking screen, and found that when participants could hear distinct action sounds such as paper ripping, the mirror neurons still fired. This showed that mirror neurons do not necessarily respond to visual stimuli specifically, but rather the understanding that usually comes with visual stimuli.

Finally, (Nelissen et al., 2005) performed a study showing that mirror neurons do not fire to biological agents only. They set up a study where stimuli consisted of video clips of several objects performing the actions 1. humans 2. just human hands, and 3. robotic hands. The results showed that the mirror neurons fired in all conditions demonstrating that an artificial stimulus such as a video is sufficient for mirror neuron activation, the entire actor's body is not required for mirror neuron activation since only seeing a hand was enough, and most importantly, non-biological agents, such as robotic limbs, and potentially also emojis, activate mirror neurons.

Face-to face-communication is multimodal since it consists of at least the auditory and the visual modalities. While auditory information comprises speech and is considered most dominant in message transmission, the visual modality which includes inter alia head movements, facial expressions and hand gestures, can disambiguate the speech content, emphasize it, change its meaning or substitute it (Goldin-Meadow, 1999; Kelly et al., 1999; McNeill, 2005; Kendon, 2004). Due to how the mirror neurons function and the type of stimuli that activates them, seeing people communicate is an inherent part of our face-to-face communication system and the visual modality is extremely important in comprehension.

There is no doubt, however, that speech or text are very powerful in transmitting messages, which is why they have been used in numerous semantic priming studies. Semantic priming studies are useful for studying semantic processing because a priming word or sentence can activate the brain in a way where the response to a probe, which is subsequently presented, will be facilitated with a faster reaction time, when the probe is related instead of unrelated. An example of this would be *cat -tiger* being processed faster than *napkin-lion* (Meyer and Schvaneveldt, 1971; Neely, 1977).

Another type of priming is affective priming. This works by presenting a priming stimulus with either a positive or negative affective valence and then measuring the behavioral and psychological response to a related or unrelated probe thereafter. The way the affective priming works is if the probe is affectively congruent to the prime instead of incongruent. This was demonstrated using emojis and just words as primes(Comesaña et al., 2013). More importantly, this task used masked primes, meaning that a distraction was presented right before the prime so that the participants were unaware of the emoji prime.

Not only did the priming effect occur even though the stimuli were covertly processed, the results show that the priming effect occurred more significantly for the emoji than it did for the words. (Comesaña et al., 2013) conclude that the results occurred due to the automatic processing of the saliency of the facial expressions being more significant that the words.

Both of these types of priming can be measured physiologically through EEG. This is done through event-related potentials, which are amplitude deflections in the EEG waveform that are related to a specific event. One ERP in particular is called the N400, which is a negative deflection approximately 400 ms after the onset of the stimulus (Kutas and Hillyard, 1980; Kelly et al., 2004; Wu and Coulson, 2005; Wu and Coulson, 2007; Holle and Gunter, 2007; Özyürek et al., 2007; Ousterhout, 2015b; Kutas and Federmeier, 2011). The way it is measured and useful, is that when a probe is incongruous to the prime, instead of congruous, the amplitude is much more negative.

There are several types of EEG recording devises ranging in the number of electrodes used for medical and consumer purposes. One consumer grade EEG system in particular is called the Emotiv headset, which utilizes 14 EEG channels. Although this system is much simpler than standard medical grade EEG devices, there have been a number of studies that support its efficacy in ERP research (Ekanayake, 2010; Badcock et al., 2013; Boutani and Ohsuga, 2013; Badcock et al., 2015; Mayaud et al., 2013; Ousterhout, 2015a; Ousterhout, 2015b; Kawala-Janik et al., 2015; Ousterhout and Dyrholm, 2013).

According to the mirroring theories and other theories which address the importance of gestures in face-to-face communication (Kendon, 2004; McNeill, 1992), multimodality is an essential aspect of the way in which humans communicate. Communicating by written texts involves the visual modality only and all the discourse content is expressed by words. Short messaging is often a replacement for oral communication, it is quick and it can therefore be difficult to express one's personality, affective state, irony, or emphasis in it. This is why emojis are becoming so popular.

This study aims to investigate the use of moving emojis in different positions of short sentences to see whether they aid or supplement text adding elements usually expressed by body behavior. A preliminary survey was conducted online in which short sentences had emojis placed in different locations and participants had to respond to which subject and/or object they thought the emojis were referring to. Successively, this information was used to place emojis in short sentences and test whether they produce enough semantic priming (N400 effect), when an incongruous probe stimuli is presented.

In the next section, an explanation of how the study was conducted is provided. This includes the pre-test survey, the description of the participants and stimuli, the procedure of the entire follow up experiment, and the method used for analyzing data. Then a summary of the results is given, explaining how the participants performed behaviorally as well as physiologically. Following this is a discussion of the results commenting on why the participants performed the way they did and what this means. Then there is a short conclusion discussing future work. Finally there is an appendix providing the pre-test survey and EEG test sentences.

2 Method

2.1 Pre-test: Survey

To figure out in which position in a sentence an emoji would be best used to refer to an element in that sentence, a survey was created with sentences where emojis in different locations had the potential to refer to multiple items in that sentence (see Appendix). Each of the sentences had 2 or 3 questions asking to which element the emoji was mostly related.

The survey was answered by 72 participants and show mixed results. Since there seems to be little pattern, and some examples directly contradict each other, it was decided to use the sentence examples where there was the most unanimity among participants. Therefore only examples with answers above 70% agreement were looked at. In most of these examples, the participants thought that the emojis referred mostly to the element (person, object or animal) which the emoji followed. In a minority of examples, in which the emojis preceded the subject of the sentence, the emoji was found to refer to the subject.

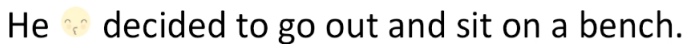


Figure 1: Sentence example with [He][was][whistling.] as congruous probe to the whistling emoji and [He][was][sad.] as incongruous to the whistling emoji.

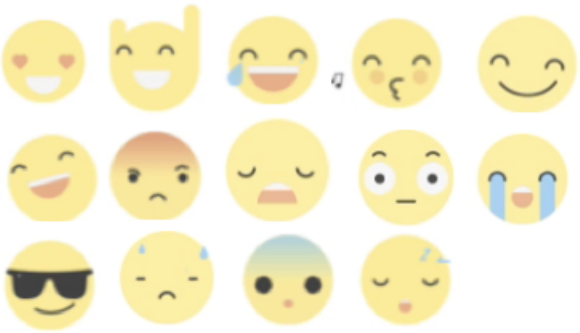


Figure 2: Pictures of all 14 moving emojis which were taken from `https://cdn.dribbble.com/users/43762/screenshots/1925708/emojis.gif`. Their meanings are as follows: loving, dancing, laughing, whistling, happy, happy, angry, sad, energetic, crying, confident, sad, scared, and sleeping.

2.2 Participants

For this experiment 19 participants were used that had an English University speaking level[1]. The mean age was 31.6 years of age with a standard deviation of 8.9 years. 10 were males, and 16 were right handed. None reported any cognitive or reading problems and everyone had good or correct-to-good vision. They all signed an informed consent document explaining that they knew what the experiment was about and that their data would be published yet individually they would remain anonymous. Due to artifacts in the EEG recordings which involved too much noise in the signal quality, 2 participants' data were eliminated.

2.3 Stimuli

The study consisted of 45 prime sentences each with a congruent and incongruent probe stimulus resulting in 90 stimuli examples in total. When the emojis are placed as verbs, the incongruent probe word is inconsistent (e.g. whistling to yelling, and laughing to crying), when the emojis are placed as adjectives the probe words are antonyms (happy to sad, and excited to bored). See Figure 1 for a sentence example and more are in the Appendix. Each prime sentence had one of 14 different moving emojis, see Figure 2 for emoji examples. These emojis would play on a replay-loop since they each only lasted 1-3 seconds. When each prime sentence was displayed, subjects had as much time as they needed to read the sentence and the experiment would continue with a button press. The sentences could be complex. Then a sequence of three slides with a single word would appear at a time, which in culmination made a phrase where the third word was always directly congruent or incongruent to the moving emoji.

2.4 Procedure

Participants were allowed to read the prime sentence for as long as was required to fully process it and would continue with a button press on a keyboard. Then there would be a fixation cross for a random duration between 750 and 1250 ms. Afterwards the three words of the probe sentence would appear in the center of the screen, each for 500 ms, and on the final probe word, which was always the 3rd word, the participant was required to make a button press response for whether the word was congruent or incongruent with the emoji in the prime sentence. This experiment consisted of 1 cycle of all 45-sentence pairs summing up to 90 stimuli sentences and taking about 20 minutes to complete. The

[1] All participants had a University degree where English was required.

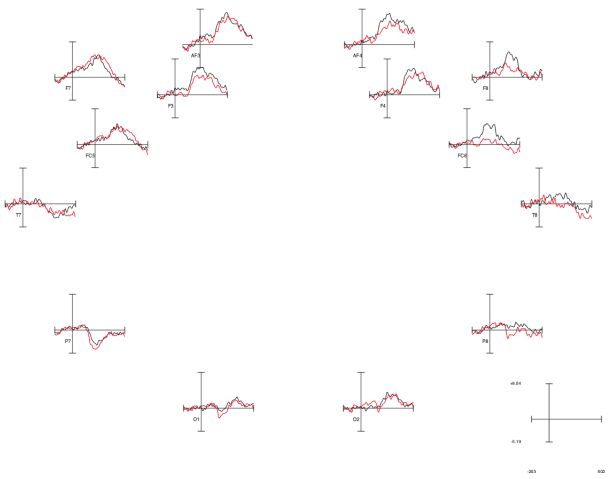


Figure 3: Waveforms of grand average responses to congruous (black) and incongruous (red) probes to emoji prime sentences.

experiment was conducted in a standard university office with potential for noise disturbances outside to support continued research with this device in more natural environments. After the experiment was concluded, participants were asked about their opinion of the experiment as a whole, if the sentences were coherent, and their thoughts about the emojis themselves.

2.5 Electrophysiological Acquisition

This investigation was conducted using the Emotiv headset for EEG acquisition which is a commercial grade system using 14 electrode channels at positions AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 and CMS/DRL references at P3 and P4 according to the International 10-20 locations. The sampling rate for the acquisition was done at 128 Hz and was filtered between 0.01-30 Hz. The artifact rejection method was done using a moving peak-to-peak system where the full width of the window was 200 ms, the window step was 50 ms, and the voltage threshold was 100 microvolts. Components were measured offline where waveforms were baselined from -200 ms to probe onset time. Waveform total duration were set from -200 to 1000 ms and were computed for analyses with Matlab's EEGLab and ERPLab, and SPSS.

3 Results

3.1 Behavioral Results

The behavioral results show an average accuracy of 87% for all congruent and 89% for incongruent responses. Also, there was an average reaction time of 1,116.4 ms for all congruent stimuli with a 1,046.8 ms response time for incongruent stimuli. A paired two-tailed t-test of these reactions times resulted in a statistically significance difference of 0.034.

Further analyses was conducted dividing the primes into two groups where the emoji referred to an object or subject before its placement in the sentence, and after it. The accuracy for before placement was 86.1% and for after it's placement it was 88.8%. The mean reaction time for before placement was 1222.38 ms where for after it's placement it was 1114.3 ms. A paired two-tailed t-test of these reactions times resulted in no statistically significance difference of 0.13.

3.2 Electrophysiological Results

Using a repeated measures ANOVA with 2 x 4 for congruency and electrode position (P7, O1, O2, P8) using a culmination of all congruent and incongruent data, there was no effect $F(1,16) = 16.0$, $P = 0.781$. See Figure 3 for the waveform and Figure 4 for the scalp map. However, performing the same repeated

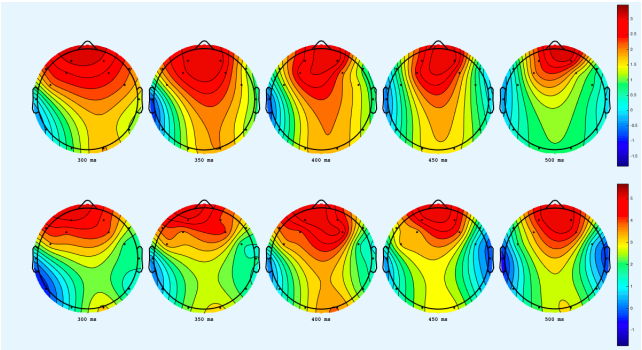


Figure 4: Scalp maps of grand average responses to congruous (top) and incongruous (bottom) probes to emoji prime sentences.

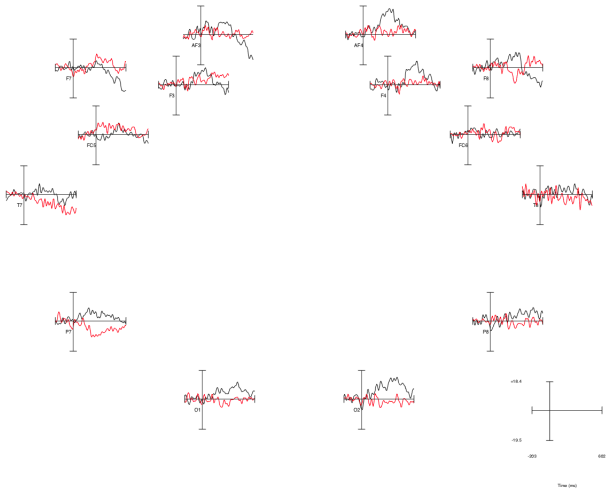


Figure 5: Waveforms of responses to congruous (black) and incongruous (red) probes to ”scared” emoji prime sentences.

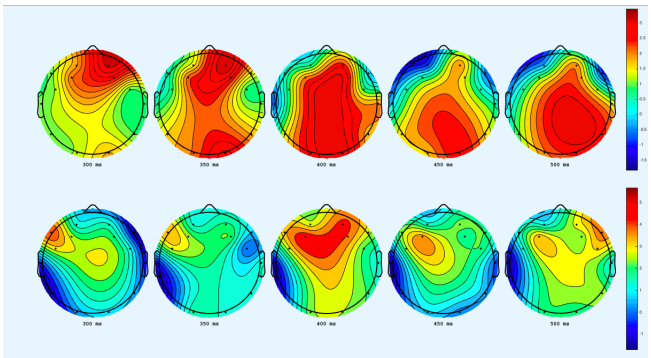


Figure 6: Scalp maps of responses to congruous (top) and incongruous (bottom) probes to ”scared” emoji prime sentences.

measures ANOVA upon analyzing the individual emojis, there was a significance for the ”scared” emoji $F(1,16) = 5.0$, $p = 0.034$. See Figure 5 for the waveforms and Figure 6 for the scalp map.

Further analyses involving the two subgroups where the emoji referred to an object or subject before it's placement or after was also conducted. A 2 (congruency) X 2 (referent direction of before/after) repeated measures ANOVA was conducted with no effect for congruency $F(1,16) = 0.276$, $p = 0.607$, and no effect for direction $F(1,16) = 0.003$, $p = 0.954$.

4 Discussion

The discovery of mirror neurons has demonstrated the importance of observation when trying to understand what others are doing (Rizzolatti et al., 1996; Rizzolatti and Fabbri-Destro, 2008; Jeannerod, 1994).This observation ability enables viewers and learners to understand why agents perform a certain action and thus allow them to predict a certain outcome (Rizzolatti et al., 1996). This mechanism is very robust and the same neural area will be activated despite if a human or monkey performs an action and also regardless of distance or proximity (Rizzolatti and Craighero, 2004). Furthermore, Freedberg and Gallese (2007) found that the mirror neurons are also fired when humans look at pictures of humans. Emojis, although not identical to human faces, are similar to them and therefore their presence can be assumed to activate mirror neurons as well and thus facilitate communication, understanding and learning. Not only that, but since mirror neurons have been shown to be activated during emotion expression (Singer et al., 2004; Wicker et al., 2003), emojis seem like an ideal method to express emotions digitally. Lastly, Nelissen et al. (2005) showed that mirror neurons fired even when a robot hand was showed on a video screen. This means the activation will occur to stimuli that do not look exactly human, and when they are very far away, and thus very small. There seems to be no limit to how artificial or abstract an emoji can be where it would be unable to transmit an intended message that could activate the mirror neuron system in the observer, which is the source for the generation of this study.

Studies such as Chu and Kita (2011) have shown how the incorporation of gesture can help with tasks such as spatial problem solving, and Broaders et al. (2007) and Goldin-Meadow et al. (2009) have demonstrated how gesturing during problems solving helps with learning through visualization techniques. Therefore, this paper attempted to show the need for co-speech gestures in text reading and how the implementation of emojis could do so. The first step was to find out where in a sentence people typically found emojis to be most descriptive when describing a specific subject. With this information, the study investigated if the implementation of various moving emojis placed strategically in sentences, would supplement readers' understanding of prime sentences by adding emotional information to the context of the sentence. To test this, semantic priming was studied through N400 production. The reason for this entire investigation was to take the first steps towards creating a universal multimodal reading system that would supplement text with images of body behavior since that is natural and crucial in our face-to-face communication.

The results of the pre-study showed that people had a very broad and varying opinion of what an emoji referred to when placed in different sentence positions. However, it seemed that most participants thought that an emoji following an element would refer to this preceding element. Therefore, using only the sentences with the most agreement among participants, which meant over 70% agreement, it was decided to place emojis directly after what they were referring to. Despite this, reaction times were slower for congruous responses than incongruous ones, and more inaccurate. Also, the grand average of all the responses to the congruous versus incongruous probes were unable to produce an N400 effect that was statistically significant. In fact, both grand averages for congruous and incongruous probes look almost identical on the scalp map and extremely similar on the waveforms. This means that the semantic supplementation of emojis was confusing on a grand average scale resulting in the same general cognitive response to both congruous and incongruous probes to the same sentences. To investigate further if any emojis in isolation produced an N400 effect, all 14 different emoji types and the sentences there in were averaged in groups and resulted in only one of them producing a statistically significant N400 effect, which was the emoji depicting fear.

The additional analyses of the behavioral and physiological reactions to the two types of sentence

probes where the emoji referred to an object or subject before it's placement in the sentence or after it resulted in no significant effects. The accuracy was lower and insignificantly slower when the emoji referred to something before it in the sentence than after meaning that participants had no real response difference regarding where the emoji was referring to in the sentence. This could mean that subjects are equally comfortable retaining the content of the sentence with the context of the emoji in the sentence at the same time on a more global level where the semantic processing effect of the subsequent probe results in no real difference. This potentially allows for greater freedom in the placement of the emoji in a sentence, as long as the reader can make sense of it.

There can be several reasons why there was no success in a kind of universal emoji complementation in sentences. The first can be seen from the pre-study, where only few sentence examples had over over 70% participant agreement on the same configuration of text and emoji. This shows that many people had extremely different opinions about where emojis should be located in sentences which could be a result of their country of origin, country they currently live in, background, age, education level, and familiarity and experience with using emojis. So even with the placement that was most agreed upon, many participants certainly still found their location to not be ideally placed. This placement problem could have caused participants to think the emojis did not refer to the subjects or objects in the sentence, but to the whole sentence holistically. Another reason for the lack of results may be due to the fact that in the internet survey the stimuli used static emojis while in the EEG experiment moving emojis were used. In the future, differences in processing the two types of emojis should be investigated. A third reason could be that the sentence primes were quite complicated and longer than in previous EEG experiments.

The lack of definitive results can be explained by many reasons, such as cultural differences of emoji experiential usage in everyday life, which is not isolated to variables of age, gender or nationality. Another explanation for the results, which was discovered during the post experiment interview process was that some subjects reported intentionally ignoring the emojis in sentences since they were under the assumption that they were there as some kind of "trick" that they did not want to be susceptible to. Without the semantic priming of the emoji in the sentence, there could neither be a congruous or incongruous probe response. This issue could not be avoided because during the participant instruction phase of the experiment, they were simply told to read the sentence as they normal would, and there was no emphasis on paying particular attention to the emoji.

Finally, there was a large discrepancy between participants regarding which emotion the emoji was supposed to portray. One notable example was that one emoji was reported to be "flirtacious" by one person while another claimed it was "angry". With such a large discrepancy on the semantic and emotional meaning and content of the emojis, there would also be a likewise disagreement in which probes were congruent and incongruent. These, along with potentially other variables account for why the grand average congruous and incongruous responses looked identical. These results are in alignment with Miller et al. (2016) who found that emojis are very open to interpretation with large variability in opinion regarding both sentiment and semantics and thus may lead to communication error. One solution to this would be some kind of standardization of emojis for particular emotions and expressions, however as Miller et al. (2016) explains, the same type of emoji is displayed differently on different company devices and platforms.

While the results of this study are inconclusive, further investigation seems important not only due to how gestures help people learn and understand, but also that when we are prohibited from gesturing, our ability to communicate becomes less fluent (Rauscher et al., 1996) and therefore, finding the best way to include emojis in text could help people express themselves properly.

5 Conclusion

This study wanted to find placements for emojis into text which would provide universal benefits in reading comprehension due to the added benefit of visual information providing body behavior which is extremely crucial for proper communication and message transmission. Following a survey which indicated that participants thought the emoji mostly referred to the element they followed in the sentence, we tested how people processed a large number of moving emojis in this position in various sentences.

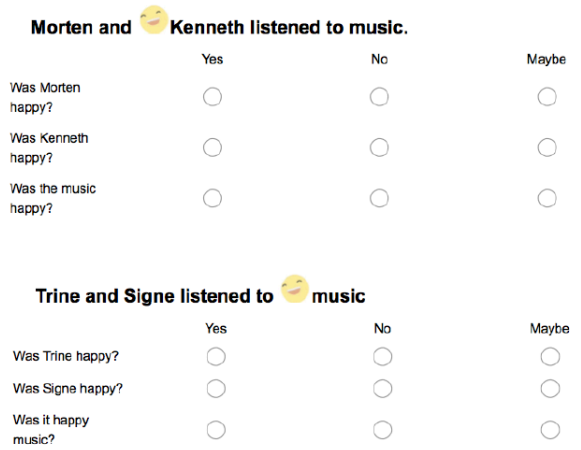


Figure 7: Examples of two questions from the survey.

Figure 8: Examples of stimuli sentences with congruous and incongruous probe sequences.

We didn't find significant differences in the N400 effect between congruent and incongruent probes, and neither in reaction time. The main reason for this was probably that the participants thought that the emojis were ambiguous and chose to ignore them when processing the prime sentence. Furthermore, the sentence primes in which the emojis occurred were quite long which might have made the task too difficult. The results also indicate that there were differences between the participants who answered the survey on the internet and the subjects who participated in the EEG experiment in terms of knowledge of infrequent emojis. Another reason for the lack of conclusive results of the EEG experiment may be due to the fact that moving emojis were used in it while static emojis were shown in the online survey. Whether there are differences in the way people process static and moving emojis should be investigated in the future. Many steps need to be taken to create such a system which involves more pilot studies regarding people's interpretations of emojis, and using a smaller target of homogeneous subjects.

6 Appendix

References

David F Armstrong, William C Stokoe, and Sherman E Wilcox. 1995. *Gesture and the nature of language*. Cambridge University Press.

Nicholas A Badcock, Petroula Mousikou, Yatin Mahajan, Peter de Lissa, Johnson Thie, and Genevieve McArthur.

2013. Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps. *PeerJ*, 1:e38.

Nicholas A Badcock, Kathryn A Preece, Bianca de Wit, Katharine Glenn, Nora Fieder, Johnson Thie, and Genevieve McArthur. 2015. Validation of the emotiv epoc eeg system for research quality auditory event-related potentials in children. *PeerJ*, 3:e907.

Hidenori Boutani and Mieko Ohsuga. 2013. Applicability of the emotiv eeg neuroheadset as a user-friendly input interface. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 1346–1349. IEEE.

Sara C Broaders, Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2007. Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, 136(4):539.

Mingyuan Chu and Sotaro Kita. 2011. The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, 140(1):102.

Montserrat Comesaña, Ana Paula Soares, Manuel Perea, Ana P Piñeiro, Isabel Fraga, and Ana Pinheiro. 2013. Erp correlates of masked affective priming with emoticons. *Computers in Human Behavior*, 29(3):588–595.

Michael C Corballis. 2002. *From hand to mouth: The origins of language*. Princeton University Press.

Hiran Ekanayake. 2010. P300 and emotiv epoc: Does emotiv epoc capture real eeg? *Web publication http://neurofeedback. visaduma. info/emotivresearch. htm*.

David Freedberg and Vittorio Gallese. 2007. Motion, emotion and empathy in esthetic experience. *Trends in cognitive sciences*, 11(5):197–203.

Susan Goldin-Meadow, Susan Wagner Cook, and Zachary A Mitchell. 2009. Gesturing gives children new ideas about math. *Psychological Science*, 20(3):267–272.

Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences*, 3(11):419–429.

Henning Holle and Thomas C Gunter. 2007. The role of iconic gestures in speech disambiguation: Erp evidence. *19*, 19(7):1175–1192.

Marc Jeannerod. 1994. The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain sciences*, 17(02):187–202.

Aleksandra Kawala-Janik, Mariusz Pelc, and Michal Podpora. 2015. Method for eeg signals pattern recognition in embedded systems. *Elektronika ir Elektrotechnika*, 21(3):3–9.

Spencer D Kelly, Dale J Barr, R Breckinridge Church, and Katheryn Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4):577–592.

Spencer D Kelly, Corinne Kravitz, and Michael Hopkins. 2004. Neural correlates of bimodal speech and gesture comprehension. *Brain and language*, 89(1):253–260.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Evelyne Kohler, Christian Keysers, M Alessandra Umilta, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297(5582):846–848.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp). *Annual review of psychology*, 62:621.

Marta Kutas and Steven A Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Peter F MacNeilage. 1998. The frame/content theory of evolution of speech production. *Behavioral and brain sciences*, 21(04):499–511.

Louis Mayaud, Marco Congedo, Aurélien Van Laghenhove, D Orlikowski, M Figère, E Azabou, and F Cheliout-Heraut. 2013. A comparison of recording modalities of p300 event-related potentials (erp) for brain-computer interface (bci) paradigm. *Neurophysiologie Clinique/Clinical Neurophysiology*, 43(4):217–227.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

David McNeill. 2005. *Gesture and thought*. University of Chicago Press.

David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. blissfully happy or ready to fight: Varying interpretations of emoji. *ICWSM16*.

James H Neely. 1977. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: general*, 106(3):226.

Koen Nelissen, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A Orban. 2005. Observing others: multiple action representation in the frontal lobe. *Science*, 310(5746):332–336.

Thomas Ousterhout and Mads Dyrholm. 2013. Cortically coupled computer vision with emotiv headset using distractor variables. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*, pages 245–250. IEEE.

Thomas Ousterhout. 2015a. Cross-form facilitation effects from simultaneous gesture/word combinations with erp analysis. In *Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on*, pages 493–497. IEEE.

Thomas Ousterhout. 2015b. N400 congruency effects from emblematic gesture probes following sentence primes. In *Intelligent Engineering Systems (INES), 2015 IEEE 19th International Conference on*, pages 411–415. IEEE.

Aslı Özyürek, Roel M Willems, Shinichi Kita, and Peter Hagoort. 2007. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Cognitive Neuroscience, Journal of*, 19(4):605–616.

Frances H Rauscher, Robert M Krauss, and Yihsiu Chen. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4):226–231.

Giacomo Rizzolatti and Michael A Arbib. 1998. Language within our grasp. *Trends in neurosciences*, 21(5):188–194.

Giacomo Rizzolatti and Laila Craighero. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192.

Giacomo Rizzolatti and Maddalena Fabbri-Destro. 2008. The mirror system and its role in social cognition. *Current opinion in neurobiology*, 18(2):179–184.

Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. 1996. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141.

Tania Singer, Ben Seymour, John O'Doherty, Holger Kaube, Raymond J Dolan, and Chris D Frith. 2004. Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661):1157–1162.

Maria Alessandra Umilta, Evelyne Kohler, Vittorio Gallese, Leonardo Fogassi, Luciano Fadiga, Christian Keysers, and Giacomo Rizzolatti. 2001. I know what you are doing: A neurophysiological study. *Neuron*, 31(1):155–165.

Bruno Wicker, Christian Keysers, Jane Plailly, Jean-Pierre Royet, Vittorio Gallese, and Giacomo Rizzolatti. 2003. Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. *Neuron*, 40(3):655–664.

Ying Choon Wu and Seana Coulson. 2005. Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. *Psychophysiology*, 42(6):654–667.

Ying Choon Wu and Seana Coulson. 2007. How iconic gestures enhance communication: An erp study. *Brain and language*, 101(3):234–245.

Physicians' and patients' use of *body-oriented gestures* in primary care consultations

Jennifer Gerwing
Health Services Research Unit (HØKH)
Akershus University Hospital
Pb.1000, Lørenskog, 1478 Norway
jennifer.gerwing@gmail.com

Abstract

Research on healthcare communication has focused little on the semiotics of gesture in interaction. This paper presents an analysis of how patients and physicians use gestures in actual consultations, focusing specifically on *body-oriented gestures* (i.e., those in which a part of the body, either through indication or demonstration, plays an integral role in the speaker's meaning). Two publically-available training DVDs for general practice consultations provided 29 minutes of excerpts from actual patient-physician encounters between nine physicians and twelve patients. All gestures were located. Body-oriented gestures were analysed for their relationship to speech and function in the interactions. Results showed that 104/238 of patients' gestures and 30/178 of physicians' gestures were body-oriented. Gesture and speech conveyed complementary information suited to each modality. These gestures served a variety of functions (e.g., establishing mutual understanding, foreshadowing information that would be contributed later, providing cohesion between topics). Just as research on healthcare communication would benefit from further exploration of the semiotics of gesture use, these findings illustrate potential for basic research: healthcare interactions offer a practical arena for investigating how patients and physicians integrate gesture and speech as they discuss consequential topics such as symptom relief, diagnosis, decisions, and treatment plans.

1 Introduction and objectives

In conversation, co-speech hand gestures are "generally recognized as being linked to the activity of speaking and are often regarded as part of the speaker's total expression" (Kendon, 1980, p. 207). Speech and gesture "cooperate to express the speaker's meaning" (McNeill, 1992, p. 11) and have been characterized as *integrated messages* (Bavelas & Chovil, 2006). The study here assumes that interlocutors can use gestures as a communicative resource (Kendon, 1994) and extends the topic from *when* gestures communicate (see Hostetter, 2011) to *what is being accomplished* with gesture. Thus the overarching purpose here is to show how doctors and patients mobilize gestures with their speech while undertaking clinical activities. In clinical communication research, the semiotics of gesture use is understudied, despite abundant theory and methodological tools from basic research. Drawing on these tools, this paper presents a gesture analysis of actual videotaped primary care interactions.

Gesture studies using videotapes of actual clinical consultations have shown that patients use gestures around their body to demonstrate the position, scale, and character of their suffering, in order to provide the sense and significance of the illness and symptoms (Heath, 2002). Patients time their expressions ('cries') of pain within the frame of diagnostic activities, as a way to balance justifying the need to seek medical help with taking an analytic orientation to their own subjective experience (Heath, 1989). Physicians use gestures to convey unique content that can be absent from speech, but which could be ambiguous without being integrated with the accompanying speech (Gerwing & Dalby, 2014). These authors reported that physicians' gestures provided unambiguous indications of the

relevant body region (e.g., the liver), but their speech provided necessary information about why it was relevant (e.g., proposing that the patient agree to diagnostic tests). In an analysis of gestures about pain collected in non-clinical settings (i.e., interviews with participants), speakers conveyed information about the location and size of pain sensations in gesture and information about pain intensity, effects, duration, case, and awareness in speech (Rowbotham et al., 2012).

The objective here was to analyze a particular subgroup of gestures that could be considered particularly salient in clinical consultations, namely those during which speakers "touch, focus on, draw or sculpt forms in front of a particular part of the speaker's body" (Calbris, 2011, p. 78). While all gestures could, in a sense, be in front of a part of the speaker's body, for this analysis, we considered gestures to be body-oriented if the hand(s) served a deictic function by directing attention to a particular part of the body (e.g., the knee, an area of skin) or if the speaker mobilized his or her body to demonstrate an action (e.g., stretching, miming using an inhaler or taking a pill). These latter gestures were akin to *character-viewpoint gestures* (McNeill, 1992), where, in this case, the speaker's hands and body represent his or her own hands and body. *Body-oriented gestures* were formally operationalized as purposive movements of the hands and/or body that were synchronized with the timing and content of speech and in which the part of the body, either through indication or demonstration, was an integral part of the speaker's meaning, as conveyed by both the gesture and concomitant speech.

The research questions were the following:

(1) How prevalent were *body-oriented gestures*?

(2) How did they relate to accompanying speech?

(3) How did physicians and patients use them?

2 Methods

Source materials were two publically available training videos (Roberts, Moss, and Stenhouse, 2003; Roberts, Attwell, and Stenhouse, 2006) containing excerpts from real primary care consultations filmed in the UK (9 physicians and 12 patients). The excerpts illustrated a variety of clinical situations. The consultations were all carried out in English. The participants represented a variety of language backgrounds (including both L1 and L2 English speakers as well as English variants), ages, and gender.

All 12 excerpts were analyzed, providing approximately 29 minutes of material. Gestures and speech were annotated using ELAN (Wittenburg, Brugman, Russel, et al., 2006). The purpose of the analysis was ultimately to locate and analyse body-oriented gestures. To accomplish this, analysis proceeded through four levels, gradually filtering the total number of gestures to those that were of particular interest. The four levels of analysis were the following: (1) All gestures were located. (2) Gestures with a semantic function were distinguished from those with a beat or interactive function. (3) Semantic gestures with concrete referents were distinguished from those with abstract referents. (4) Semantic gestures with concrete referents were examined further to identify those that were oriented towards the body. The following sections provide brief operational definitions that guided these decisions.

2.1 All gestures were located: Gestures vs. adaptors

Gestures were defined as observable hand or arm movements that (1) had a rhythm matching the rhythm of speech and (2) had a gesture stroke with a purposive, clear direction or trajectory. They were distinguished from instrumental hand actions (e.g., typing on a keyboard, opening a file folder) and from self-oriented adaptors (e.g., scratching the arm). Thus similar hand movements could be classified as gesture or adaptor based on the timing the accompanying speech. For example, one patient said, "I've had this years and years and years ago but it wasn't as bad- I just got a little rash".

Timed with the speech "this years and years and years," he made scratching motions around his fingers. This movement would be considered a gesture because the onset of his gesture stroke was timed precisely with the word "this", the scratching motion was clear and almost stylistic, and the scratching movements matched the quick rhythm of his speech. In contrast, earlier in the consultation, the same patient made small, vague, fiddling motions with his fingers; these motions began while the doctor asked a question about whether he had started using any new washing powders, and they continued through the entirety of the patient's next utterance ("no not no- no- nothing like that, as I said it didn't happen- it was about seven o'clock I woke up very uncomfortable"). This fiddling motion began gradually, with no precise onset, and the movements were not timed with the patient's speech. Therefore, it was considered an adaptor, not a gesture. All gestures and concomitant speech were annotated. Each of these gesture-speech composites was analyzed for its immediate function (semantic/beat/interactive).

2.2 Gesture function: Semantic vs. Beat vs. Interactive

Semantic gestures functioned to convey topical content to the conversation, (Gerwing & Dalby, 2014) either through representation (e.g., rotating the shoulders to demonstrate shoulder stretches) or deixis (e.g., pointing to a shoulder to show the location of pain). The finger scratching motions in the gesture example above (while the patient said "I've had this for years and years and years") served a semantic function, depicting what the patient meant by "this", namely, an itchy skin condition.

Beat gestures functioned as discourse markers; these non-representational, small up and down movements served merely to emphasize the words with which they were timed (McNeill, 1992). For example, a patient explained that the pain in her hip continued when she went for walks, and the doctor checked his understanding by saying, "it's worse when you walk". During the words "worse" and "walk", he made two up and down movements with his left hand. These were considered beat gestures because they were non-representational and were timed precisely with two key words in the doctor's utterance.

Interactive gestures served a social function. Like beat gestures, they were non-representational; however, their form differed from beat gestures: the speaker would rotate his or her hand, so that the palm was displayed, and move the gesture towards the addressee. Furthermore, rather than fitting the context of speech at the level of content, these gestures served to manage interactive aspects of the conversation (e.g., Bavelas, Chovil, Lawrie, and Wade, 1992). For example, in one of the consultations, a doctor explained to the patient that her smoking was elevating her blood pressure. He said, "and I'm not just saying that". The patient replied, "yeah, yeah, it's just the truth" while nodding, holding her right palm up, and moving it towards the doctor. This gesture was interactive: firstly, because its form matched the form identified in Bavelas, et al. (1992), secondly, because it served to credit the doctor's assertion that he was contributing more than just his opinion about smoking. Once again, note that gestures with a similar form could serve different functions, depending on the context that speech supplied. For example, if the patient in the immediately previous example had accompanied her palm up movement with the speech, "the rash started here", the gesture would be considered semantic, because it was contributing the location of where her rash began.

2.3 Abstract vs. Concrete referents

Each semantic gesture was analyzed for whether it depicted a concrete or abstract referent (Gerwing & Dalby, 2014). This decision depended on the relationship between the form of the gesture and what was being conveyed in speech. For a gesture to have a *concrete referent*, its form or manner would have to represent or point to an object, person, location, or action. The above examples of semantic gestures had concrete referents: The rotating shoulders gesture referred to an advisable action; pointing to shoulders referred to a region of the body; scratching the fingers referred to the location of a skin condition and its quality. *Abstract referents* were metaphorical depictions of abstractions, concepts, or ideas. For example, one patient, explaining her feelings of sadness to the doctor, said, "it's like, back in your mind, all your life, do you know, like a film". During "like a film", she slowly rotat-

ed her hands in outward circles, as though depicting the motions of a film reel. This gesture illustrated the image of a turning film reel, showing the experiences of her life in her mind. This gesture was thus not depicting an actual film (in which case the gesture would be considered as having a concrete referent); it illustrated a metaphorical film and therefore had an abstract referent.

2.4 Body-oriented gestures

Finally, concrete semantic gestures were further examined to identify those that were oriented towards the body. Note that the above analytical process filtered all gestures such that body-oriented gestures were defined as purposive movements of the hands and/or body that were synchronized with the timing and content of speech and that made a region of the body an integral part of the speaker's meaning. Body-oriented gestures could include pointing towards the speaker's own body (e.g., a patient could indicate where she experienced pain) or the addressee's (e.g., a physician could point to an area on the patient's body to ask whether it was the locus of pain). Demonstrations of actions included postural portrayals of feelings (e.g., sitting up straight suddenly to portray shock). For a concrete semantic gesture to be body-oriented, the speaker's body had to be representing a body: For example, if the speaker pumped his arms to mimic walking, the gesture would be body-oriented. However, if the speaker demonstrated walking by "stepping" with his fingers across the table, it would not be body-oriented. Body-oriented gestures were selected for detailed qualitative analysis to explore their functions in the consultations.

3 Results

In the 29 minutes of material, there were 416 gestures. Patients gestured at a rate of 11.42 gestures per 100 words; physicians at 6.37 gestures per 100 words. Table 1 provides a differentiation among the types of gestures patients and physicians used, reported as raw frequencies.

Table 1. Number and functions of patients' and physicians' gestures.

	Patient	Physician
Semantic concrete- body oriented	104	30
Semantic concrete- not body oriented	22	17
Semantic abstract	49	69
Beat	33	50
Interactive	30	12
TOTAL	238	178

3.1 Patients' body-oriented gestures

For patients, 104 of their 238 gestures were body-oriented. In these, gestured information complemented but was rarely redundant with information conveyed in the speech. Gestures indicated the relevant body part (e.g., the chest), demonstrated body positions, indicated direction within the body (e.g., the radiation of pain), or depicted objects or scenarios around the patient (e.g., drawing a doorway, using the hand as a telephone). Speech conveyed the sensation (e.g., pain, lack of pain, tenderness, itchiness), intensity, temporal aspects (e.g., whether the patient experienced a symptom in the past or present, repetitiveness, duration of symptoms), the depth of the relevant region (e.g., skin, muscles, bones), and the circumstances around injuries or illnesses. Of the 104 body-oriented gestures, 66 were direct references to a particular body part (as opposed to a demonstration of an action). Of these direct references, in 49 (0.74), the patient did not specify the body part in the speech. Sometimes the name of the body region was missing from speech entirely, with speech instead indicating deictically that the gesture would be providing information (e.g., to describe where he had a rash, one patient said "it's not just here, it's here, here[1]…" while pointing to his armpits, thighs, and groin). Some-

[1] In this and other examples, underlining indicates the timing of the gesture with speech.

times, patients mentioned a general body region in their speech, but narrowed the focus by specifying the precise location in gesture (e.g., one patient said she had "pain on her leg", while pointing to her left knee and thigh).

3.2 Physicians' body-oriented gestures

For physicians, 30 of their 178 gestures were body-oriented. These gestures anchored their questions or explanations. Physicians rarely named the body part or region when they gestured. Of the 30 body-oriented gestures, 24 were direct references to a particular body part (as opposed to a demonstration of an action), and in 21 (0.88) of these, the physician did not specify the name of the body part in the speech. For example, one physician pointed to an area on his chest while saying "pressing on these muscles here these bones". Another, initiating the clinical examination, said "let me have a look at the skin" while pointing to the patient's arm.

3.3 Examples of how patients and physicians used body-oriented gestures

Physicians and patients used body-oriented gestures *to display their understanding to each other*. For example, in response to a physician's explanation of how to apply lotion after his baths, one patient nodded and made patting motions around his torso (where he had patches of dry skin). Earlier in the consultation, the physician had examined the patient's skin, thus the location of the dry skin was already established. Rather than functioning to re-orient the physician's attention to those areas of his chest, these patting gestures displayed the patient's understanding of the physician's instructions, contributing to accomplishing the clinical task of securing mutual understanding regarding his treatment plan. Physicians and patients used body-oriented gestures to sort out misunderstandings. One sequence is presented in Figure 1, which illustrates three gestures in a question-answer sequence about chest pain. First, the physician used his right hand to point to a location on his chest while asking for elaboration about the patient's symptoms. His speech ("can you tell me about this pain, does anything make it worse?") did not specify where the pain was, but his gesture did, displaying where he had understood the patient to be experiencing pain (see Figure 1A). The patient then corrected him by indicating on her own body where she saw him pointing (while saying "it's not pain there"; Figure 1B) and then slightly lower, where she actually felt the pain (while saying "it's uh- it's inside me there"; Figure 1C). Thus she indicated the location of her symptom, foregrounded against the physician's displayed understanding. Once they had established the location of her symptoms, she could answer his question regarding whether anything made the pain worse.

Figure 1. Physician and patient use body-oriented gestures to clarify the location of the patient's pain.

Patients sometimes used body-oriented gestures *to foreshadow information before they articulated it clearly*. For example, some patients would touch the location of a symptom once or twice before beginning to speak about it. In one consultation, a patient described the circumstances of an injury she had experienced at work. At the beginning of her description, she performed a gesture in which she brought her hands together in front of her chest (see Figure 2, frames A and B). At first, she accompanied the gesture with the speech "the alarm, the alarm", which did not make its referent clear. In the second gesture, she added "fire" and "the doors". Finally, she added a more dramatic movement, start-

ing the gesture with her hands at either side of her (Figure 2C) and bringing them together quickly (Figure 2D), while saying "and close". The speech and gesture here, together, demonstrated doors closing directly in front of her, something that was only foreshadowed in the two previous gestures. At that point, the physician (who had been watching intently) nodded and responded, "right, right". The patient continued, pointing at the consequent injury on her ankle, which then became the focus of her and the physician's gaze (see Figure 3).

Figure 2. The patient uses body-oriented gestures to foreshadow the circumstances of her injury.

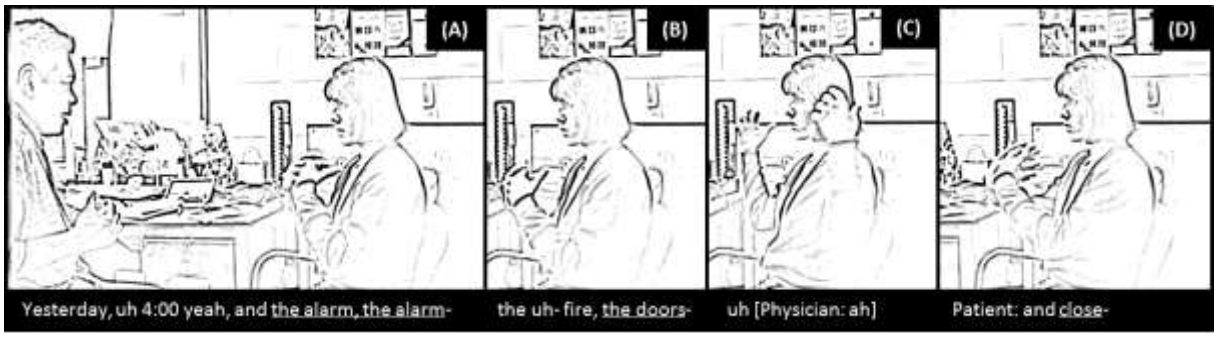


Figure 3. Patient exhibits her injured ankle to the physician, which they both look at.

Patients used body-oriented gestures *to put information that might be considered unpalatable into gesture* rather than speech. For example, two patients used a gesture that moved from the throat up to and out of the mouth to convey vomiting, although neither put that information in their speech. Figure 4 presents a frame-by-frame series showing this gesture; the patient's accompanying speech was "and I was very bad, with terrible pain, with sickness you know". Later, the physician asked, "When was the last time you actually were sick, when you, you vomited?" demonstrating that he had indeed understood her gesture, even though she had said nothing in her speech about vomiting.

Figure 4. Sequence of screen shots showing patient's gesture conveying vomiting.

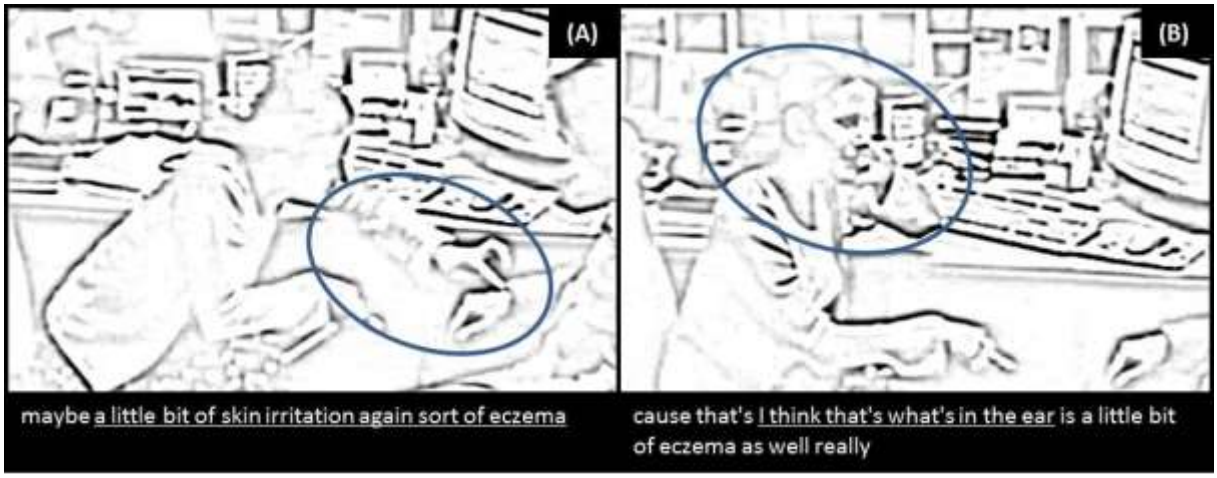
Patients used body-oriented gestures with speech *to contrast past and present emotions or attitudes*. For example, one patient described the terrible side effects of an antibiotic. While saying "after the fourth day, oh I dreaded it I took them all, but I dreaded it", she drooped her body over the table (as if in dread) while smiling. Figure 5 shows this position. This latter, complex portrayal did more than locate symptoms; instead, the patient used her body and speech to provide an evocative demon-

stration of her response to her body's reaction to the antibiotics, showing both despair in her posture (which her speech placed in the past with "took them all" and "dreaded it") and normalcy and even good humour with her facial display.

Figure 5. Patient's simultaneous portrayal of past dread and present amusement.

Physicians repeated body-oriented gestures *to connect topics during the consultation*. For example, one physician explained to the patient that the itching and pain in her ear was caused by irritation in her ear canal. The physician gestured towards her own left ear during the explanation, and described the patient's ear canal as being "inflamed" and "red". Less than a minute later, she examined the patient's hand, which also had a rash on it. The physician leaned forward and used her left index finger to draw circles over the back of the patient's hand while saying "maybe a little bit of skin irritation again sort of *eczema*" (see Figure 6A), thus introducing the topic of eczema. She then connected the eczema on the patient's hand to the previously discussed irritation in the patient's ear, saying "cause that's I think that's what's in the ear is a little bit of eczema as well, really" while motioning to her own ear again (see Figure 6B). Her gesture to her own ear echoed her previous gesture; repeating this body-oriented gesture provided visible cohesion to her earlier explanation of the patient's ear irritation, reinforcing that the symptoms might be connected by a systemic condition.

Figure 6. A physician provides cohesion between symptoms by repeating an earlier gesture.



Physicians also used gestures towards the patient's body *to demonstrate their clinical activities*. The physician above, while examining the skin irritation on the back of the patient's outstretched hand, used her thumb and index finger to frame the area. This gesture at that moment served no purpose except to demonstrate that she was examining that specific location.

4 Discussion

Physicians and patients integrated body-oriented gestures with their speech in sophisticated and systematic ways. By using gestures to inquire about and show locations of symptoms, physicians and patients referred unambiguously to relevant body regions without resorting to potentially ambiguous terminology. These gestures accomplished far more than simply conveying information about symptom location; both patients and physicians used them to perform complex clinical communication tasks. Further research can focus on how health care providers and patients integrate gesture and speech in a variety of clinical settings (e.g., ones involving language barriers, electronic medical records, emergent and non-emergent scenarios). Besides generating a new appreciation for the integral role visible action plays in clinical communication, such research could have implications for healthcare systems, as they explore new media for interaction (e.g., telehealth, telephone interpreters).

This study illustrates how researchers accustomed to investigating the gestures interlocutors produce in experimental, laboratory conditions need not feel restricted from studying them in everyday settings. As long as both interlocutors are captured on video, how they use co-speech gestures can be explored systematically and in detail. Healthcare interactions thus offer a practical arena for basic research on gesture use. Patients and health care providers discuss consequential, high-stakes topics such as symptom relief, diagnosis, prognosis, decisions, and treatment plans; systematicity in how they integrate gesture and speech during these discussions could shed light on the role gesture plays in dialogue more widely. Further research using the analytical framework described here is being conducted on interactions between specialists and patients in a Norwegian hospital and on triadic, interpreted primary care encounters filmed in the UK.

References

Bavelas, Janet Beavin, Nicole Chovil, Douglas A. Lawrie, and Allan Wade. (1992). Interactive gestures. *Discourse Processes, 15*, 469-489.

Bavelas, Janet Beavin and Nicole Chovil. (2000). Visible acts of meaning. An integrated message model of language use in face-to-face dialogue. *Journal of Language and Social Psychology, 19*, 163-194.

Calbris, Genevieve. (2011) Elements of meaning in gesture. The Netherlands: John Benjamins B.V.

Gerwing, Jennifer and Anne Marie Landmark Dalby. (2014) Gestures convey content: An exploration of the semantic functions of physicians' gestures. *Patient Education and Counseling, 96*, 308-314.

Heath, Christian. (1989). Pain talk: The expression of suffering in the medical consultation. *Social Psychology Quarterly*, 113-125.

Heath, Christian. (2002). Demonstrative suffering: The gestural (re) embodiment of symptoms. *Journal of Communication, 52*(3), 597-616.

Hostetter, Autumn B. (2011). When do gestures communicate? A meta-analysis. *Psychological bulletin, 137*(2), 297.

Kendon, Adam. (1980) Gesticulation and speech: two aspects of the process of utterance. In: Key MR, editor. The relationship of verbal and nonverbal communication. The Hague: Mouton Publishers. p. 207–27.

Kendon, Adam. (1994). Do gestures communicate? A review. *Research on language and social interaction, 27*(3), 175-200.

McNeill, David. (1992) Hand and mind. What gestures reveal about thought. Chicago: University of Chicago Press.

Roberts, Celia, Becky Moss, and J. Stenhouse. (2003) ' 'Doing the Lambeth Talk': Patients with Limited English and Doctors in General Practice: an educational resource' DVD and handbook. Centre for Lan-

guage, Discourse and Communication, King's College London. London: NHS London Post-graduate Deanery.

Roberts, Celia, Christine Atwell, and Stenhouse, J. (2006) 'Words in Action : an educational resource for doctors new to UK general practice'. London: NHS London Post-graduate Deanery.

Rowbotham, Samantha, Judith Holler, Donna Lloyd, and Alison Wearden. (2012). How do we communicate about pain? A systematic analysis of the semantic contribution of co-speech gestures in pain-focused conversations. *Journal of Nonverbal Behavior*, *36*(1), 1-21.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. (2006) ELAN: A professional framework for multimodality research. LREC 2006, fifth international conference on language resources and evaluation.

Multimodal health communication in two cultures – A comparison of Swedish and Malaysian Youtube videos

Jens Allwood
SCCIIL Center
Univ or Gothenburg
jens@ling.gu.se

Elisabeth Ahlsén
SCCIIL Center
Univ of Gothenburg
eliza@ling.gu.se

Stefano Lanzini
SCCIIL Center
Univ of Gothenburg
lanzhbk@hotmail.it

Ali Attaran
Univ of Malaya
Kuala Lumpur
ali.attaran.ts@gmail.com

Abstract

Youtube video health information about overweight and obesity, was analyzed in two different countries – Sweden and Malaysia. The videos were analyzed by using Activity based Communication Analysis, Critical Discourse Analysis and Rhetorical Analysis, pointing to possible cultural differences in rhetorical approach. The use of multimodality was in focus in the analysis. Considerable differences in the use of spoken and written words, pictures, animations, colour, music and other sounds were found between Swedish videos which tended to rely more on spoken words from experts and on logos, while Malaysian videos tended to rely heavily on animations, vivid colours, music and other sounds and appeal to pathos. In both countries, ethos is important, but conveyed in somewhat different ways. The length of the videos differ considerably, with Malaysian videos being very short and Swedish videos quite long.

1 Introduction

The paper presents a comparative study of multimodal communication in Youtube videos on obesity and health from Sweden and Malaysia. The main message in the videos is to inform people about health and obesity, in order to make them want to lose weight as well as to inform about and/or sell ways of doing this. The study, as part of a more comprehensive project on communication about overweight and obesity, compares the rhetoric of multimodal videos on overweight and obesity produced in the two countries. The purpose of the study is to address the question of whether health communication needs to be adapted to different cultures or not.

The ancient traditions of rhetoric in Western countries, as established by Aristotle and in Eastern countries as inspired by Confucius, were to some extent different, as they were developed in different contexts and these differences can still be noticed in different areas of rhetoric and argumentation. The main focus in the Western tradition is placed on rhetorical logos (reasoning), while in the Eastern tradition, relatively more emphasis is placed on pathos (i.e. appeal to the reader's or listener's emotions), due to the importance of establishing and maintaining good personal relations, and a combination of logos and pathos is often advocated (Zhu and Hildebrandt, 2003, Wang, 2006)). The Western and Eastern division seems to be applicable to Sweden and Malaysia, although it must be stressed that there is considerable variation within both Western and Eastern geographical areas and cultures.

Other cultural differences, often ascribed to Eastern and Western cultures are greater individualism in more low-context Western cultures and greater collectivism in more high-context Eastern cultures (Hall, 1976, Hofstede, 1991).

Another cultural difference between Sweden and Malaysia is the potential target group, i.e., the population and their demographic features. One obvious difference is the religious and ethnic characteristics of the population. Sweden is largely and traditionally a protestant Christian country

-

with a mainly European population and with a tradition of social welfare and relatively high mean level of education, although lately more varied, due to immigration. Malaysia has a multicultural and multireligious tradition, with a main Malay muslim population, but also very noticeable Chinese and Indian groups. The level of education has been lower than in Sweden, but has lately shown a considerable increase. Some comparable figures from the two countries, provided by UNESCO, are that the gross enrollment in primary education is 84% in Malaysia and 96% in Sweden, while the proportion of pupils starting grade 1 who reach grade 5 is 99.2% in Malaysia and 100% in Sweden. The literacy level in Malaysia is 98% for persons between 15 and 24 years of age. The secondary school enrollment in Malaysia is about 70%. When we turn to higher (post secondary) education, the figure for Malaysia was 21.6% in 2010 (noting an increase of about 5% per 10 years). In Sweden, 42% have entered tertiary education and about 40% hold a tertiary degree (sources: UNICEF, 2017). There is, thus, still a certain, although diminishing, difference in mean educational level.

The analysis used three different framworks for analysis, Activity based Communication Analysis (ACA) (Allwood, 2013), Critical Discourse Analysis (CDA) (Fairclough, 1995) and Rhetorical Analysis (Kennedy, 2007)

2 Method

2.1 Material

Video films on Youtube, containing information and or propaganda about overweight and obesity were identified for Sweden and Malaysia. Five Malaysian and four Swedish videos were chosen as representative, after scanning available Youtube videos providing information and recommendations concerning overweight and obesity to the public. Observers rated the available videos for typicality among the available Youtube videos and the videos obtaining the highest total scores in each of the countries were selected for analysis. Tables 1 and 2 provide short descriptions, data on time, persons appearing and means of expression used in each video.

Table 1. Swedish Youtube videos

Name of video	Short description	Time	Persons appearing	Means of expression used:
An increasing number of children are overweight”	A local TV news clip about overweight and fat children working out in a new ”fun” program, involving a game	2 min, 03.02 sec	Children taking part in program activity 2 children interviewed 1 organizer/expert interviewed	Speech Text Filmed activity/context
Fighting against obesity	A commercial for clinic with a specific operation technique for obesity	3 min, 13.28 sec	1 expert presenting Pictures Patient seeing doctor	Speech Text Pictures Filmed activity/context
Overweight	An information video about health and obesity from a health website with a famous doctor	4 min 30.14 sec	People walking in street and at beach Expert talking in street and at beach Doctor and patient Person cooking	Speech Text Pictures Film Music
Weight line	A video from a health website, with a famous doctor promoting a specific diet	1 min, 46.15 sec	Expert in white coat at desk Patient and doctor Expert in nature People training	Speech Text Pictures Filmed activity/context Music

Table 2. Malaysian Youtube videos

Malaysian videos

Name of video	Short description	Time	Persons appearing	Means of expression used:
The balloon	A short film of a boy blowing up a balloon, then of many persons blowing up balloons and balloons bursting/exploding	30.00 sec	1 boy, then 12 persons blowing up balloons	Sound Text Filmed event
Information film from the Ministry of Health/The fat man	A fat man having hearth pain - dramatic. A voice speaks for him and tells how much he has been eating. Then about changing lifestyle and the man looks happy.	32.02 sec	1 fat man	Speech (speaker voice representing person in film) Filmed events Texted warning
Obesity in Malaysia	Texts about obesity and a healthy lifestyle are mixed with pictures of a happy family and contact data and commercial for slimming product.	2 min, 13.21 sec	Pictures of families of 4 persons	Speech (speaker voice) Text Pictures Film?
Malaysian obesity		1 min, 14.51 sec	No persons	Sound Music Pictures Text

The data in the table are further discussed below.

2.2 Overview analysis using Activity based Communication Analysis, Critical Discourse Analysis and Rhetorical Analysis

Three frameworks, Activity based Communication Analysis (ACA) (Allwood, 2013), Discourse Analysis (CDA) (Fairclough, 1995) and Rhetorical analysis (Kennedy, 2007) were used. The ACA and CDA analysis were selected to give an overview description of what activities were shown and what the purpose of the producer might be for each video. The rhetorical analysis is the main method and is used more in-depth for analyzing the approaches used in Sweden and Malaysia, respectively.

The analysis noted what type of expressive means occurred in each of the videos. Since both verbal and nonverbal modes of expression are used and several modes can be used at the same time, and, furthermore, different rhetorical modes can be used simultaneously, the study mainly uses qualitative content analysis, rather than quantitative comparison. However, the analysis contains observation of features that were very salient and/or dominating in the videos. Speech and text were analyzed as *verbal communication,* whereas all other means of expression used in the videos were, for the purposes of this paper, considered as *non-verbal*.

2.2.1 Activity based Communication Analysis

An overview of factors determining the social activity was made for each video. These factors were the purpose of the activity, the social roles of the activity, the instruments used and the environment (cf. Allwood, 2013).

ACA is a framework which relates analysis of behavior, in the form of production and perception of communicative expressions and interactive patterns in social activities to an analysis of influencing background factors of the particular activity at hand.

The influencing background factors first of all consist of the collective conditions linked to the social activity, such as its purpose, inherent roles, physical and biological context and psychological/social context. But they also consist of individual background factors carried into the activity by the participants and linked to their background and characteristics as well as the individual purpose, role, physical/biological conditions and psychological/social factors of the activity in the eyes of each of the participants, depending on their background.

The background factors are considered essential for understanding what happens in communicative behavior.

The part of ACA used in the analysis of this study is the analysis of influencing background factors for each film, according to the characteristics:

Social Activity
Collective Background factors of the Video:
Purpose
Roles in communication
Physical/biological factors
Psychological/social factors
Background factors of the participants:
Purpose
Roles

The social activities (as well as individuals involved) are complex and can be embedded in the analyzed videos, i.e. there can be a producer behind a film as well as main spokesperson in the film and there can be an audience or recipient of information in the film as well as an assumed audience of viewers of the Youtube videos.

2.2.2 Critical Discourse Analysis

For each video, the main question of "in whose interest" the video was made was addressed (cf. Fairclough, 2005). Each video film has a more ore less explicitly and unambiguously stated sender and a purpose, as noted above, and the CDA analysis strives to make explicit both overtly expressed and assumed interests of the sender. The senders, in this case, are the producers and/or sponsors of the video, sometimes appearing in the video as persons, sometimes in logos or texts, sometimes not at all. They can also include other agents expressing the message in the video. The audience is important for the message and it is therefore also of interest to find out what the target group is for the message of the video.

The main question asked in the present analysis was: In whose interest was this video made and who benefits from it?

2.4 Rhetorical analysis

The main analysis was a rhetorical analysis of multimodal logos, ethos and pathos. Logos, ethos and pathos were identified and described for each of the videos in each country/culture and then compared between Sweden and Malaysia.

Logos refers to the content of the argument presented, its premises and conclusions, the internal consistency, clarity and type of the claims that are made, the type and strength of the supporting evidence (Aristotle, in Kennedy, 2007). The impact of logos on an audience is sometimes called the argument's logical appeal (Ramage et al., 2015).

Ethos refers to the trustworthiness or credibility of the writer or speaker (Aristotle, in Kennedy, 2007). The impact of ethos is often called the argument's "appeal from credibility" (Ramage et al., 2015)

Pathos refers to persuading by appealing to the reader's/ listener's emotions, attitudes and/or imagination. This can be called the argument's emotional influence or appeal (Aristotle, in Kennedy, 2007). Appeal to attitudes and emotions can be approached from two perspectives: *expressive* (the emotions and attitudes expressed by the speaker/writer) and *evocative* (the emotions and attitudes the speaker/writer is trying to evoke in the reader/co-communicator) (Allwood, 1978).

The rhetorical analysis was made according to the categories of the schema in figure 1.

	Verbal	Non-verbal
Logos		
Pathos		
Ethos		

Figure 1. Coding categories for intended functions of expressive means in the Youtube videos.

The occurrence and form of each category were given short descriptions in the table for each of the videos.

3 Results

3.1 Analysis of an example of Swedish Youtube videos – The Weight Line (Health care video)

This video is from Sweden, with Swedish speaking senders and recipients and the topic is obesity information.

Activity based Communication Analysis:

The purpose is 1. to inform the public about health risks connected with obesity and 2. To give information to obese and overweight people about what they could do in order to lose weight. There are three roles in the video: 1. A male expert giving information – a well known professor, 2. A female doctor, 3. A female patient, 4. The intended viewers are the general public, especially obese and overweight people. The instruments are an internet website, text, pictures and films. The environment is a health care network giving obesity informtion.

Critical Discourse Analysis - In whose interest is this video made?:

The video is in the interest of obese and overweight persons, who can get more information about the possible risks and ways of their weight and how to lose weight. The expert professor can get more visibility and increase his credibility. All in the category of physicians and dieticians benefit by creating good will for themselves. The health website The Weight Line and the professor benefit from showing their brands during the video and two companies show their logos in the video, promoting these companies.

Rhetorical analysis:

Ethos - Verbal: The expert uses statistics and medical terms and the text and presents his job position and workplace. Overlay websites URLs appear during the video. An overweight woman talks about her way of losing weight.

Ethos - Nonverbal: The expert professor and the female doctor are wearing white doctor's coats. The expert professor is a well-known and recognized person and the female doctor is sitting in a medical room. The person who needs help is clearly overweight.

Pathos – Verbal: The participants use vocal and written words describing diseases. These can evoke concern and fear. The words pronounced by the overweight woman "For this time I will succeed" try to evoke inspiration and hope for overweight viewers.

Pathos – Nonverbal: The woman talking about her way to fight obesity can be seen to be overweight, promoting inspiration and hope in overweight viewers.

Logos – Verbal: The expert professor explains the health risks of being overweight and obese. He explains the "4M" factors to fight againts overweight and obesity. The overweight woman presents how she is following the suggestions of the professor. The female doctor suggests to the audience that they read more information on the "Weightline" website and invites overweight people to talk to their physician.

Logos – Nonverbal: Video images of food and of people doing physical excercise are shown.

3.2 Analysis of an example of Malaysian Youtube videos – The Bursting Balloon

This video is a short fim – The bursting Balloon, warning about the risks of obesity. The film is from Malaysia, aimed at English speaking recipients and contains obesity information.

Activity based Communication Analysis:

The purpose is to inform the public about health risks connected with obesity. The roles are: 1. Writer, producer (signed or anonymous), 2. The characters in the film, 3. The intended viewers. The instruments are an internet website, mainly animated film and short text. The environment is an official website giving obesity information.

Critical Discourse Analysis – In whose interest is this video made?:

The video is made in the interest of obese and overweight persons, who can be alerted about the possible risks of over-weight and obesity. This is also in the interest of the Malaysian health care system and Malaysian tax payers.

Rhetorical analysis:

Logos – Verbal: Only one sentence, "You'll end up like this balloon, if you don't control your diet.", containing an implication, but mostly appealing to pathos (se below).

Logos – Nonverbal: Balloons are being blown up and explode, the eplosion of the balloon being a metaphor of the risks connected with obesity, increase of obesity ending in catastrophe

Ethos – Verbal: The video has text showing that it comes from a credible source, an official government website.

Pathos: Verbal – The video has only the one final written message (in red) "You'll end up like this balloon, if you don't control your diet.", intended to capture attention, create fear of obesity and desire to lose weight.

Pathos – Nonverbal: The increasing noise of the sound from blowing up the balloons, the increasing size of the balloons and the bursting of the balloons create a feeling of increasing threat and danger, also creating fear of obesity.

3.3 Comparing the rhetoric of the Swedish and Malaysian Youtube videos

***Swedish logos*:** is expressed by experts (using medical terms) and obese persons explaining the problems of overweight and consequences and recommending solutions, like talking to a doctor, exercising, eating less, eating a specific diet, consulting a website, contacting an organization and explaining benefits of loosing weight in speech and text. The logos arguments are *nonverbally* supported by the appearance of experts and obese people, graphic diagrams, images of food and working out and or fat

people interacting with doctors. This makes the Swedish videos dependent on long sequences of speech.

Malaysian logos: consists of very short verbal reports on increasing obesity in Malaysia, why people get fat, i.e. through fat food, fast food availability and use of cars, and the consequences of obesity. Logos is not so important – there is more stress on pathos.

Swedish ethos: In the *Swedish* videos, a famous physician expert on obesity, and nurses are talking or being interviewed, with their names, titles, workplaces in overlay text. Medical terms are used. Obese people describe their problem and/or their successful treatment This is non-verbally supported by the physician and nurses wearing white coats or suit and tie, by the logos of authorities or clinics, by the environments of hospital, medical clinic, office, book shelves and a gym, and by the appearance of obese people telling about problems and treatments.

Malaysian ethos is also achieved by reference to health authorities in a text that is a video produced by an authority or by a well-known gym chain, that a product is linked to a Harvard professor, and by first hand experience of obese persons. It also contains a quote from a famous person, an appeal to patriotism and an imperative tone (from the health authority). In two of the videos, there is no identifiable appeal to ethos. In other videos, there is nonverbal support through the appearance of the logo of the gym and by obese persons telling their story. In general, there is much less focus on experts in the Malaysian videos.

Swedish pathos: Swedish Youtube videos show a happy, healthy family and gym activity as well as overweight persons talking about having fun while training, having found the right diet etc. This can evoke sympathy and inspiration. There is also talk about problems, risks, and diseases, which can evoke fear, and talk comparing treatment methods, where the one being promoted is described as new, and widely accepted in other countries, possibly evoking a feeling of safety. Nonverbally, pathos is achieved by showing obese people standing on scales, and people have serious faces, possibly evoking unpleasant feelings or fear. Overweight persons shown having fun while training and shown talking about having found right diet, on the other hand, evoke sympathy, giving inspiration. This is further supported by relaxing music at the end of the videos, to evoke positive feelings.

Malaysian pathos: The verbal part consists of the use of frightening words like: *terrifying, threat, impending doom* (videos 1 & 5), and *severe* (in bold) (video 4), as well as a warning slogan - warning (in red font). We can see that the text is presented using also nonverbal features (bold and red). In addition, there is music creating fear (like a rising pitch), image (a bursting balloon). This is combined with a text warning in red font in a video without speech, relying heavily on pathos. There is also music creating fear and then calm and upbeat music when possible solutions are introduced. Music is used in all Malaysian videos. In addition, in one video the setting is dark grey, the actor's dress dull, the facial expressions showing suffering.

4 Discussion and conclusions

Even though there are some similarities, for example in the use of ethos, conveying ethos by relying on authoritative sources and making use of self-experienced obesity in both countries. However, major differences also seem to be at work in the two countries, some of the main points of interest concerning:

1) The length and type of videos and the choice of expressive media – words, film and music. In Malaysia, more films and music is used and the message is often conveyed through metaphor or metonymy, whereas in Sweden, more words and factual pictures are used.

2) The use of logos, pathos and ethos. Logos is used more in Swedish videos, in the form of facts and explanations. In the Malaysian videos, logos is represented by frightening descriptions. Pathos, on the other hand, is used more in the Malaysian videos. In the Swedish videos, suggestions are made to the viewers, but in the Malaysian videos, there are more commands and emotional action evocation. The use of status-based ethos (expertise, reputation) is more similar, although it occurs slightly more in the Swedish videos. In both countries ethos is also conveyed through self-experience of the persons needing help.

There could be many factors behind the differences found between Swedish and Malaysian videos about obesity and overweight. Why are short and very multimodal videos with an "advertising" approach, designed to capture attention common in Malaysia, but not found at all in Sweden? Why are long videos with elaborate verbal explanations of facts common in Sweden, but do not seem to occur in Malaysia? One possible explanation is (i) different cultural traditions of expression in general, for example more use of vivid colours in Malaysia. (ii) It could also be related to the intended audience, where the Malaysian audience comes from more varied ethnic-cultural backgrounds and have more varied levels of education than the Swedish one. (iii) A third explanation is that Malaysia is more traditionally authoritarian then present-day Sweden with more use of "stick" than "carrot". A similarity between the countries is that ethos is important in both. In Malaysia, ethos is achieved by showing the name and logo of a ministry, while in Sweden, the focus is more on individual experts, presenting their titles and institutions, using statistics and other scientific findings and presenting them in a professional (hospital) environment. This could reflect a more absolute and accepted authority for government agencies in Malaysia and more reliance on experts and science results in Sweden. This could also account for the more direct use of imperatives in Malaysian videos and statements making up longer explanations and suggestions in the Swedish videos.

The take home message of our article, is that there seem to be both verbal and non-verbal cultural rhetorical differences in the way health information about obesity is provided and that global health information, in order to be effective should probably take such differences into account.

References

Allwood, J. (1978). *On the analysis of communicative action*: University of Gothenburg.,
Allwood, J. (2013). A multidimensional activity based approach to communication. In Wachsmuth, I., de Ruiter, J., Jaecks, P. & Kopp, S. (eds) *Alignment in Communication*. Amsterdam: John Benjamins, pp- 33-55.
Aristotle, In Kennedy, G. A. (2007). *On rhetoric: a theory of civic discourse*. New York: Oxford University Press.
Fairclough, N. (1995). *Critical Discourse Analysis - The Critical Study of Language*. London: Longman.
Hall, E. T. (1976). *Beyond Culture*. Anchor Books.
Hofstede, G. (1991). *Cultures and organizations : software of the mind*. London: McGraw-Hill.
Ramage, J. D., Bean, J. C., & Johnson, J. (2015). *Writing arguments: A rhetoric with readings*: Longman.
UNICEF (2017). statistical information: Education,
https://www.unicef.org/infobycountry/sweden_statistics.html (retrieved 20170617)
https://www.uniceforg/infobycountry/malaysia_statistics.html (retrieved 20170617)
Wang, B. (2004). A survey of research in Asian rhetoric. *Rhetoric Review*, 23(2):171-81.
Zhu, Y. & Hildebrandt, H. (2003). Greek and Chinese classical rhetoric: the root of cultural differences in business and marketing communication. *Asia Pacific Journal of Marketing and Logistics*, 15(2):89-114.

I am definitely certain of this!
Towards a multimodal repertoire of signals communicating a high degree of certainty

Laura Vincze
Dipartimento di Filosofia, Comunicazione e Spettacolo - Università Roma Tre
laura.vincze@gmail.com

Isabella Poggi
Dipartimento di Filosofia, Comunicazione e Spettacolo – Università Roma Tre
isabella.poggi@uniroma3.it

Abstract

When talking to other people, speakers do not only communicate their beliefs but also their degree of commitment towards such beliefs, their "epistemic stance", and they do so by means of both verbal and body markers.
The paper presents an analysis of the body markers that speakers use to convey high certainty and obviousness of the beliefs communicated while exposing detailed knowledge falling into their professional expertise to an audience of peer hearers. In a corpus of 42 video abstracts where doctors and medical researchers orally illustrate their scientific findings, two signals of high certainty (*headshake* and *eye-closure*) and two of obviousness (*Palm Up Open Hand* and *shoulder shrug*) are analysed from a semantic and cognitive point of view. The differences and relationships among these body markers and their verbal concomitant markers, along with their semantic nuances, are illustrated in depth.

1. Introduction

In everyday interaction speakers are invited, often even compelled, to display their knowledge on a different range of topics. Yet, given the important function of communication as exchange of beliefs, one does not only have to provide the information one knows, as required by the cooperative principle (Grice 1957), but also to make it clear how confident one is in stating something. Such meta-information concerning the beliefs conveyed has been studied in various research domains, and called "degree of certainty" in cognitive models of communication (Castelfranchi & Poggi, 1998; Poggi, 2007), and "epistemic stance" in Linguistics (Dendale and Tasmowski 2001; De Haan 2001; Kärkkäinen 2003; Cornillie 2010; Marín Arrese 2011; Zuczkowski et al., 2017).

Epistemic stance has been defined as the degree of commitment (or confidence) of the speaker towards the validity of the communicated information. If the speaker's commitment towards the truth of the information to be communicated is medium or low, the speaker adopts an uncertain epistemic stance; if such commitment is high, the speaker's epistemic stance is certain. Epistemic stance hence concerns the ways speakers comment on the possible accuracy or credibility of a claim, the extent they want to commit themselves to it (Hyland 1998). As Hyland (1998) notes, speakers must calculate what weight to attribute to an assertion, perhaps claiming protection in the event of its eventual overthrow.

Speakers and writers can hence decide to withhold complete commitment to a proposition, allowing information to be presented as an opinion rather than an accredited fact. A low degree of commitment can be communicated by means of morphosyntactic or lexical resources: for instance by "hedges" like *possible, might* and *perhaps*. Hedges, therefore, imply that a statement is based on plausible reasoning rather than certain knowledge, indicating the degree of confidence it is prudent to credit to it (Hyland & Tse 2004).

On the opposite pole stand "boosters", i.e. lexical markers of the speaker's absolute certainty, communicating a speaker's high level of certainty and commitment towards the communicated belief. Boosters allow writers to express their certainty in what they say and to mark involvement with the topic and solidarity with their audience (Hyland 2005). They underline the writer's conviction in his argument, or stress shared information and group membership.

The role of hedging and boosting is well documented in academic writing as communicative strategies for conveying reliability and strategically manipulating the strength of commitment in order to achieve interpersonal goals (Hyland 2005).

Along with *verbally* expressing their high commitment towards the communicated belief, speakers deploy a variety of *body* and *voice signals* to convey their level of certainty and possibly boost their social image as knowledgeable and authoritative speakers.

Although the encoding of epistemic stance has constituted a widely debated topic in linguistics, traditional studies in this field have focused predominantly on how speakers use *lexical* and *morphosyntactic features* to convey their commitment towards the communicated belief (Dendale and Tasmowski 2001; De Haan 2001; Kärkkäinen 2003; Conrad & Biber 2000; Hoye 2008; Cornillie 2010; Marín Arrese 2011; Zuczkowski et al., 2017, among others). Far fewer studies have adopted a multimodal perspective where verbal resources are integrated with *voice, facial and body signals* for a *multimodal* analysis of a speaker's overall commitment (Dijkstra et al. 2006; Debras & Cienki 2012; Mondada 2013; Roseano et al. 2014; Ricci-Bitti et al. 2014; Jehoul et al. 2017).

The present study aims to widen the traditional perspective adopted by works in the field of epistemicity by investigating speakers' *multimodal communication of epistemic stance*. Namely, our particular concern here is with the multimodal communication of speakers' *certainty*. More precisely, the goal of our work is to analyse the signals of certainty performed through body behaviours, namely hands, face and shoulders.

2. Boosting one's certainty by means of words and gestures

So far research in linguistics has mainly focused on the communication of *uncertainty*, and gave less attention to the communication of the *certain* epistemic stance. The reason for this must be the fact that the certain stance is typically conveyed by unmarked declarative sentences (Simon-Vandenbergen & Aijmer 2007; DeLancey 2001). Studies on Italian speakers demonstrated that unmarked declaratives suffice to convey information considered certain (Bongelli & Zuczkowski 2008), while certainty adverbs are generally used not only to convey the speaker's degree of commitment to his beliefs, but have to do with the ways speakers want to position themselves in current discourse, for instance as engaged and expert speakers (White 2003; Simon-Vandenbergen & Aijmer 2007).

For example it is shared knowledge that the politicians' role is to persuade their audience, and since to express knowledgeability and certainty is part of their roles as persuaders, the expression of knowledgeability and certainty can indirectly be a cue to social identity or authoritativeness (Simon-Vandenbergen 1996). A speaker may emphasize his certainty and authoritativeness by using high certainty markers (boosters) like *clearly, obviously, demonstrate,* but also multimodal markers, whose role is to underline the speaker's conviction in his argument and boost speaker's reliability.

The communication of epistemic stance (both certain and uncertain) in health context has so far been mostly focused on *written texts* (Salager-Meyer 1994; del Olmo 2014). *Oral communication* was generally investigated in *asymmetric* contexts like doctor-patient interactions (Peräkylä 1997; Douglas & Heritage 2005; Landmark et al. 2015). To our knowledge, so far no study has approached doctors' multimodal communication in a *symmetric* context. The present paper analyses doctors' and medical researchers' multimodal communication of certainty and authoritativeness while orally illustrating their findings to a public of *peers*.

3. Corpus and Method

To investigate body signals of certainty, in a corpus of 100 video-abstracts where medical researchers orally illustrate their findings published in the British Medical Journal[1], with the aim of a more rapid dissemination of their research to peers, we selected 40 videos (for a total of 186 minutes) in which authors speak freely in front of the camera, without reading from a script. The speakers are gender-balanced (62 speakers, 28 females and 34 males) and, with the exception of 3 males and 3 females, all native English Speakers.

Our aim was to find out to what extent upper body movements, hand and facial gestures (i.e. *shoulder shrugs*, *palm up open hands*, *headshakes* and *eyelid closure*) play a communicative role in conveying speaker's high commitment towards the communicated beliefs. Body signals informing on the speaker's

1 The British Medical Journal invites authors of research papers to record a short video abstract (4-5 min in average) for publication alongside their papers. Video abstracts enable authors to explain their research findings in person, increasing the reach and understanding of their work. The videos are published on bmj.com with their articles and on multimedia (www.bmj.com/multimedia) and YouTube (www.youtube.com/user/BMJmedia) channels.

certainty towards the communicated beliefs were hence singled out and analysed. The video corpus was annotated for each participant's *shoulder shrugs*, *palm up open hands*, *headshakes*, and *eyelid closures* lasting longer than a *blink*; this was done with sound off to avoid possible biases based on the content of co-occurring speech. The videos were then viewed with audio on, to check whether the selected body signals actually bear those meanings, with the help of co-occurrent speech and interactional context. Each signal was described as to its parameters of shape and movement, its verbal context and co-occurring body communication was annotated, taking note of the combination between two or more concomitant signals; then the meaning assumed in that context was considered. Finally, the signals were grouped in terms of their meaning: certainty and obviousness.

4. High certainty vs. Obviousness

Within the signals by which the Speaker displays high certainty, we may distinguish two classes:

1. signals of high certainty, by which the Speaker claims higher epistemic status than the Hearer, since the information presented falls more in his territory of information (Kamio 1994) – i.e. the speaker is more certain than the hearer about the truth validity of such information
2. signals of obviousness: the information presented presumably falls at the same degree into both Speaker's and Hearer's territory of information, and the Speaker in some sense makes appeal to the Hearer's acknowledgement.

In the following we list the body markers of high certainty and obviousness.

5. High certainty markers

Speakers tend to communicate their high certainty concerning the delivered information by means of verbal or body signals that we call *high certainty* markers. They employ such markers when they assume that their statements may raise surprise or doubt in the hearer, and they have the goal of persuading the hearer of the truthfulness of their statements, as well as of their own reliability and trustworthiness as sources of information.

Since a different epistemic status between speaker and hearer is presupposed in this case, here the emphasis is on the speaker's commitment to the stated information, on his epistemic status and, in general, on the speaker as an authoritative source whom the hearer is invited to trust.

5.1 The "intensity headshake"

A first signal communicating a speaker's high degree of certainty is the *headshake*, defined as *rotating one's head horizontally*, either *to the left or the right, and back again, one or more times*, the head always returning finally to the position it was in at the start of the movement (Kendon 2002).

So far *headshakes* have not been seen in terms of epistemic markers but as signals conveying negation (Ekman & Friesen 1969; Kendon 2002; Robinson & Heritage 2015) or alternatively as indexing intensification and inclusivity (Goodwin 1980; McClave 2000).

Headshakes assume a function of **intensification** when used in association with positive evaluative statements (such as *'what a marvellous sunset'* – to use Kendon's 2002 example). This association between head shakes and "intensified" positive assessments has been previously noted by Schegloff (1987), Goodwin (1980), and McClave (2000), the last two scholars suggesting that an intensification headshake should be treated differently from "negation" headshakes and rather be regarded as an "assessment marker". Kendon (2002) instead claims that even the intensity headshake gives expression to a negative: if one says *"what a marvellous sunset!"*, the implied negative is: "No sunset is more marvellous".

While we totally agree with Kendon's view, we argue that this type of "intensity headshake" can be seen as an intensifier of the speaker's degree of commitment to the asserted belief. In other words, shaking one's head while stating *"What a marvellous sunset"*, does not only imply that "no sunset is more marvellous", but in our opinion also conveys the speaker's total commitment to this belief.

The "intensity" *headshake* that displays high certainty can have two different meanings, of *quantity* and of *intensification*, revealed by the concomitant words. Throughout our corpus of videoabstracts, we found a total of 15 instances of *headshakes* co-occurring with both **quantifiers** (such as *many; a lot of; dozens; several; over 80%; a wide range of*) and **intensifiers** (such as *very, particularly, by far, best*).

Let us see some examples of *headshakes* accompanied by quantifiers:

(1) *There is **many many**[2] clinical trials done on medications so people should stick to their medications but not think that that's enough 'cause what we have found is you can do so much more […].*

(2) *[…] and I think that it has **a lot of** ehm applicability*.

All the co-occurent quantifiers communicate a large quantity of something and the speaker's belief that "it is a lot". In fact, according to Johnson (1987) the likely basis for the *headshake* co-occurring with quantifiers is the speaker's referring to a collection of things taking up more than one point in space. This is why, as proposed by Birdwhistell (1966), such *headshakes* may work as a *pluralization marker*: a high quantity of something can be conceptualized as taking up more space, and such ampler space translates into an ample lateral rotation of the head.

Yet, in our corpus *headshakes* also co-occur with **intensifiers** such as *very*, *particularly*, *by far* or *best*. *Very* and *particularly* add the modified adverb or adjective a meaning of "stronger" or "more intense"; *best* is a superlative, while *by far* makes a superlative adjective even stronger.

(3) *Age is **by far** the most important risk factor for zoster.*

We argue that the very fact of intensifying the adjectives in one's sentence may work as a signal of a speaker's high commitment towards the truth value of the stated belief, probably because the high intensity of some phenomenon does not leave room for any doubt about its existence. For instance, in example (3), along with the message that age is the most important risk factor for zoster, the speaker also conveys her total commitment to this belief. Both the *headshake* and the lexical item *by far* (a synonym of *undoubtedly*) are indicators of epistemic strenght. As Kendon (2002) would put it, in saying *"Age is by far the most important risk factor for zoster"*, the implied negative here is: "There is no risk factor that is more important". Like in the sunset example, the *head shake* makes reference to that implied negation, but also conveys the speaker's high certainty in saying so.

According to McClave (2000), the broader concept of **inclusivity** is needed to account for the lateral sweeps co-occurring with lexical items such as *whole* (or *any* and *every* which we encountered in our corpus in addition to *whole*). As McClave puts it, "to intensify is to add more of something, thus increasing it in energy, volume, or number. More of something is conceptualized as taking up more space, at least metaphorically, so in this way the concept of intensification is related to the concept of inclusivity" (McClave 2000:7).

These forms of intensification and absolute inclusivity communicate a speaker's high certainty in the communicated belief, possibly aiming at persuasive goals through inducing increasing certainty in the listener as well. As shown, for instance, in the overuse of superlatives and intensification in charismatic leaders (Poggi & D'Errico, 2016), this might depend on the fact that stating a high level of something may contribute, in many cases, to communicate a high commitment towards the stated belief.

Our point is that starting from the original meaning of the *headshake* as an intensifier of events in the world, there is a shift in meaning towards intensifying a belief on the Speaker's mind (Poggi 2007). More precisely, from meaning "it is a lot", the *headshake* starts to work as an epistemic indicator of the speaker's high degree of certainty, finally meaning "it is very much so". By *shaking his head*, the sender communicates that he is highly committed to the stated belief, and possibly does so to persuade the addressee of the truth validity of such belief.

5.2. Eye-closure

The same meanings conveyed by *headshakes* (**quantitative intensifier** and **epistemic indicator of speaker's high degree of certainty**), can be conveyed by another body signal: *eye-closure*. The two body signals, *headshake* and *eye-closure*, can occur either concomitantly or independently.

Vincze & Poggi (2011) proved that some closings of the eyes (*winks* and *eyeclosure*) during speech are communicative, i.e. carry meaning. The *eye-closure* is a closing of both upper and lower eyelid, lasting longer than a physiological *blink* (defined as a *rapid closing of the eyelids and return to eyes open*, simply aimed at keeping the standard humidity of the eye); a *wink* is a unilateral closing of eyelids.

Blinks and *eye-closures* differ in duration and tension: *blinks* are brief, *eye-closures* are longer. *Eye closures* are often characterized by tension (or pressure) in the eyelids.

2 In all the cited examples, the words concomitant to the body signal under analysis are highlighted in bold.

These gaze signals may occur both in the speaker and in the listener: by means of an *eye closure* the speaker conveys the strenght of his commitment to his own belief, while the listener, producing a *headnod* accompanied by *closed eyelids*, conveys his strong agreement with the speaker's beliefs. Accompanying one's *nodding* or *headshaking* (as a signal of either negation or intensification), by an *eye-closure* conveys a higher degree of commitment with respect to the *nodding* or *headshaking* alone (Vincze & Poggi 2011). Such eyelid signal may be paraphrased as "Absolutely, I am very certain of this". The *eye-closure* hence adds to the meaning of "yes" for the nod (or "no" for the negation *headshake*) an element of "categoricity", i.e., a high level of certainty and commitment to one's beliefs.

Same as *headshakes*, *eye closures* occur with both quantifiers (in our corpus, with *many, three million, widely)* and intensifiers (*even more*).

In the examples below *eye-closures* co-occur with quantifiers.

(4) *Electronic health records are linked by **many** countries for medical research.*

(5) *We used a research data set from the Calibre programme which links four sources of health information on over **three million people** in England.*

In both cases, the speakers *close their eyes* for longer than a *blink* while referring to big quantities (*many countries* and *over three million people*). While the *sweeping of the head* in *shakes* might iconically represent the speaker's referring to a collection of things taking up more than one point in space (from where the necessity of turning one's head to visualize them all in a row), the *eye-closure* while mentioning numerically challenging groups, might evoke the speaker's closing his eyes in order to mentally picture such high quantities (large quantities need more time to be visualized).

Same as *headshakes*, *eye closures* can also work as epistemic indicators of speaker commitment, like in the following example where the speaker signals her strong commitment to the belief that linking records with other data sources is highly valuable for research by means of a long *eye-closure*.

(6) *[…] the recent linkage in general practice records with other data sources means that they are **even more** valuable for research.*

To sum up, quite in a parallel fashion, *headshakes* and *eye-closures* co-occurring with quantifiers convey the **meaning** of **high quantities** ("it is a lot"**)**, while *headshakes* and *eye-closures* co-occurring with **intensifiers** function as epistemic indicators of the **speaker's high degree of certainty** ("it is very much so", "it is absolutely so"). Throughout our corpus of videoabstracts, we found a total of 5 instances of eyeclosures, 2 co-occuring with intensifiers such as *even* and *very* and 3 with quantifiers such as *many, widely* and *three millions*.

6. Obviousness markers

Sometimes a speaker communicates to the listener that the information he is delivering is not only certain, but in a sense, over-certain, definitely obvious. In other words, not only is the speaker himself certain of that belief, but he is also certain that the listener too is already in possession of that information. While a piece of information presented as highly certain falls more into the speaker's territory of information than into the listener's (i.e. the speaker is entitled to be more certain than the listener about the truth validity of such information), information presented as obvious presumably falls at the same degree into both speaker's and listener's territory of information: speaker and listener have equal epistemic status.

As pointed out by Goodwin (1979), Sacks (1992), Stivers et al. (2011), interactants generally have good command over who can be expected to know what. Although the BMJ videos are addressed to a large audience of hearers that speakers are not personally acquainted to, the fact that hearers are peers (hence assumed to hold similar epistemic status as the speakers), allows speakers to safely assume what listeners are expected to know. As a matter of fact, in our corpus speakers often present their statements as being **obvious**, undebatable, unquestionable, hence already known to both speaker and hearer. Interestingly enough, acknowledging the hearer already knows about some issue is useful to the speaker in backing his own point of view (I acknowledge you because you must undoubtedly think the same as me).

6.1. The Shoulder shrug: "it is obviously so"

The *shoulder shrug* is a bodily booster communicating obviousness – hence implicitly speaker's high degree of commitment towards his own statements. This signal – *shoulders first raised and then going back down to their initial position* – is a polysemous item that may assume, in different contexts, quite diverse meanings, such as: obviousness; ignorance (lack of knowledge); and non intervention (either because of carelessness or powerlessness) (Jokinen and Allwood, 2010; Debras and Cienki, 2012; Debras, 2015).

For the purpose of our study we will only look at the first type of *shoulder shrug*, the one conveying the obviousness of a state of affairs. In line with Debras and Cienki (2012), we believe epistemic shrugs can work both as an *individual epistemic stance marker* (boosting speaker's own degree of commitment towards the stated belief – corresponding to our **high certainty markers**) and as an *intersubjective epistemic marker* (marking the shared knowledge between speaker and listener, or between speaker, listener and the rest of the world – both corresponding to our ' **obviousness markers**). Here are some examples extracted from our BMJ corpus.

6.1.1. The shoulder shrug as an individual epistemic marker of *high certainty*

We encountered cases of shoulder shrugs that convey speaker's high certainty when the speaker talks about the results of her/his studies. Here we have a speaker who conducted six pooled studies with over 3500 women and found that screening in healthcare settings doubled identification of partner violence, particularly in antenatal settings.
The speaker goes on saying that:

(7) *Overall, this represents **an absolute benefit** of forty-three per thousand screened women*.

Her *shoulder shrug* can be interpreted as a booster of speaker's high certainty that such result is beneficial.

6.1.2. The shoulder shrug as an intersubjective epistemic marker

Shoulder shrugs are usually intersubjective markers indicating shared knowledge either between the speaker and the listener, or between speaker, listener and rest of the world. The latter type of shrug refers to a doxic knowledge (common belief, popular opinion), and might be paraphrased as "everybody knows it". Like in the following example, where a doctor talks about safe criteria to diagnose miscarriage.

(8) *What our data shows is that if you do that [bring people back for a scan], you are possibly going to have a false positive diagnosis, in a small number of cases, we are talking about a small number or cases. But remember, you know, **there should be no errors** over something as important as this*.

What the doctor says is that when scanning women twice for miscarriage, it is possible, in a small number of cases, to have a false positive diagnosis, i.e. to diagnose a miscarriage when there has actually not been one. While sending the warning *"but remember, **there should be no errors** over something as important as this"*, the speaker *shrugs his shoulders* and *smiles*. The *shoulder shrug* conveys the obviousness of his statement, while the *smile* can be attributed to having realized that his warning was so obvious, predictable and expected that it was unnecessary to even mention it: it works as a self-ridiculization, meaning something like "I feel stupid even mentioning it".

In other cases, the information occurring concomitantly with a *shoulder shrug* falls in the territory of information of speaker-and-listener working in the same field (professional expertise). In the following example, the speaker *shrugs her shoulders* while stating that old patients are more at risk of harmful effects of breast cancer treatment, hence implying that such information is also shared by her peer listeners.

(9) *Breast cancer screening has been under debate for many years because despite the fact that it may save lives due to early detection, it may also result in a proportion of over diagnosis and over treatment and this is especially so in older patients because it has been shown that old patients are at increased risk of **harmful effects of treatment***.

The meaning of the *shoulder shrug* (whether of obviousness, ignorance or carelesness) can be disambiguated either by the parameters of the *shrug* itself (e.g. an *asymmetric shrug* – one shoulder only – generally conveys speaker's carelessness and non implication on the issue talked about), or by other

concomitant signals such as the *Palm Up Open Hand* gesture (*PUOH*), (Müller 2004; Kendon 2004; Streeck 2009). According to Streeck (2009), who considers *shrugs* as "compound enactments" with *PUOH* as a component part, a prototypical *shrug* involves several body parts: raised eyebrows, the hands turned so that the palms face up; the forearms generally lifted, and raised shoulders.

Whether we consider the *PUOH* as a gesture on its own, as Müller (2004) and Müller & Cienki (2008) do, or a component of the *shoulder shrug*, as Streeck (2009) does, the *Palm Up Open Hand* often accompanies the shrug in communicating obviousness. Throughout our corpus of videoabstracts, we found a total of 14 instances of shoulder shrugs, of which 1 accompanied by a PUOH gesture.

6.2. Palm Up Open Hand (PUOH[3])

The *PUOH* gesture, which has a configuration of *palm open*, *fingers extended* more or less *loosely*, *palm turned upwards*, often used with a *downward movement or turn of the wrist and a hold* in the end, can have two different meanings. Speakers from different cultures use *PUOH* to communicate either obviousness or lack of knowledge. According to Müller (2004), such opposite meanings derive from two domains of action: (1) giving, showing or offering an object by presenting it on the open hand and (2) displaying an empty hand, where the empty hand indicates the fact of not having something and openness to the reception of an object. A core difference between the two domains of actions is that in the first one the hand is full, i.e. a metaphoric object (like a fact or a speaker's thought) is lying there on the open hand, whereas in the second one the displayed hand is empty.

We will focus on the first meaning conveyed by the *PUOH* gesture, which thanks to the metaphor of a hand whose palm is open and visible to everyone, typically conveys that what the Speaker is saying is so evident as to be obvious.

> (10) *Breast cancer screening in general has been under debate for many years because despite the fact that* ***it may save lives*** *due to early detection, it may also result in a proportion of over diagnosis and over treatment.*

While saying *it may save lives* (due to early detection), the Speaker performs a *Palm Up Open Hand* gesture (*PUOH*), communicating that she considers such information as given and highly certain for both herself and the audience of peers.

The *PUOH* can occur either independently (as in ex. 10), or simultaneously with another body marker conveying obviousness: the *shoulder shrug*. In ex. (11), the same speaker from example (10) explains what a successful breast cancer screening programme implies, namely that the detection of early stage tumour increases, while as a consequence of that, the incidence rate of advanced tumours decreases. She performs a *PUOH* gesture concomitant to a *shoulder shrug* while saying:

> (11) *So* ***this is actually a key condition*** *for a successful screening programme*.

The two body signals convey the obviousness of her statement: what a successful screening programme should do is detecting tumours while still at an early stage.

7. Conclusion

This paper has presented some work aimed at outlining a repertoire of signals which, in verbal and body modalities, convey information on how certain the speaker is of the beliefs s/he is delivering. Two main classes of signals have been overviewed: signals of **high certainty** proper, that is, epistemic boosters, when the beliefs at issue fall in the ***speaker's*** territory of information; and signals of **obviousness**, when they fall in both interactants' territory, and the speaker in a sense makes appeal to knowledge shared with the listener and hence taken for granted. In our opinion, both types of signals – of *high certainty* and *obviousness* – convey similar degrees of speaker certainty, what differs is *how speakers envisage the hearer's stance*. Namely, high certainty markers are used when speakers tend to expect hearer's

3 Unfortunately, in most cases the videorecording focused on speakers' upper part of the trunk, leaving out the hands. This limitation of our corpus is the reason why we found only 2 occurrences of the PUOH gesture.

disagreement, surprise or *doubt* and they want to persuade him/her of the correctness or truthfulness of their point of view; while obviousness markers are used when speakers tend to expect *hearer agreement* upon the statement at issue.

Moreover, as already mentioned, in some cases the very act of presenting something as obvious may have a rhetorical value, in that it may be used by the speaker in a fallacious or manipulatory way: s/he pretends something to be obvious to avoid the burden of proof and to skip more accurate investigation on the part of the listener, as well as and to exhibit a high (fake) self-confidence. If this is the case, our hypothesis is that the same signals of obviousness might be used both in this last, manipulatory case, and in the sincere case, i.e., when the speaker really believes the topic or argument at hand is obvious. Yet, further study is needed to test this hypothesis.

The signals we have overviewed, namely uses of *headshakes* and *eye-closure* to signal high certainty and uses of *Palm Up Open Hand* and *shoulder shrug* to signal obviousness, as already mentioned, do not convey only the meanings illustrated here, but they are polysemous signals, often bearing different, sometimes even opposite meanings. Subsequent work will analyse the whole polysemy of such signals, to set out their semantic differences and the common cores of meaning, or their cognitive links. In the same vein, future work will examine the multimodal combinations of signals that Speakers use to convey their confidence in the information they are communicating.

References:

Aikhenvald, Alexandra (2004). *Evidentiality*. Oxford, Oxford University Press.ird

Birdwhistell, Ray (1966) Some relations between American kinesics and spoken American English. In: A. Smith, ed., Communication and culture, 182-189. New York, Holt, Rinehart and Winston.

Bongelli, Ramona, and Andrzej Zuczkowski (2008) Indicatori linguistici percettivi e cognitivi. Aracne, Roma

Castelfranchi, Cristiano, and Isabella Poggi (1998) *Bugie, finzioni e sotterfugi. Per una scienza dell'inganno.* Carocci, Roma

Cienki, Alan, and Cornelia Müller (2008). Metaphor, gesture, and thought. In Raymond W.Gibbs (Ed.) *The Cambridge handbook of metaphor and thought*, pp. 483-501.

Cornillie, Bert (2010). An interactional approach to epistemic and evidential adverbs. In Gabriele Diewald and Elena Smirnova (Eds.), *Linguistic Realization of Evidentiality in European Languages*. Berlin / New York, de Gruyter, pp. 309-330.

del Olmo, Sonia Oliver (2014) Hedging and attitude markers in Spanish and English scientific medical writing. *Communicating Certainty and Uncertainty in Medical, Supportive and Scientific Contexts* 25: 273-290.

Debras, Camille, and Alan Cienki (2012). Some uses of head tilts and shoulder shrugs during human interaction, and their relation to stancetaking. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. IEEE, 2012.

Debras, Camille (2015). Visual stance markers: is shrugging lexical or grammatical? Oral presentation at 13th International Cognitive Linguistics Conference, Newcastle. Theme session: Grammar, Speaker's gestures, and Conceptualization.

Maynard, Douglas W., and John Heritage (2005). Conversation analysis, doctor–patient interaction and medical communication. *Medical education* 39.4: 428-435.

DeLancey, Scott. (2001) The mirative and evidentiality. *Journal of pragmatics* 33.3: 369-382.

Dendale, Patrick, and Liliane Tasmowski (2001). Introduction: Evidentiality and related notions. *Journal of pragmatics* 33.3 (2001): 339-348

Dijkstra, Christel, Emiel Krahmer, and Marc Swerts (2006). Manipulating uncertainty. The contribution of different audiovisual prosodic cues to the perception of confidence. In R. Hoffmann and H. Mixdoff (eds.), *Proceedings of the Third International Conference on Speech Prosody*.

Ekman, Paul, and Wallace V. Friesen (1969) The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1.1: 49-98.

de Haan, Ferdinand (2001). The relation between modality and evidentiality. In Müller, Reimar & Reis, Marga (eds.), *Modalität und Modalverben im Deutschen.* Linguistische Berichte, Sonderheft 9. Hamburg: H. Buske, 201-216.

Hyland, Ken (1998). *Hedging in Scientific Research Articles*. John Benjamins.

Hyland, Ken, and Polly Tse (2004) Metadiscourse in Academic Writing: A Reappraisal. *Applied Linguistics* 25 (2): 156-177.

Hyland, Ken (2005) Stance and engagement: A model of interaction in academic discourse. *Discourse studies* 7.2: 173-192.

Jehoul, Annelies, Geert Brône, and Kurt Feyaerts (Forth.). The shrug as marker of obviousness. Corpus evidence from Dutch face-to-face conversations. *Linguistics Vanguard*.

Johnson, Mark (1987) The body in the mind: The bodily basis of meaning, imagination and reason. Chicago, IL: University of Chicago Press.

Jokinen, Kristiina, and Jens Allwood (2010). Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging. In Toru Ishida (Ed.) *Culture and computing*. Springer Berlin Heidelberg, 2010. 55-70.

Kamio, Akio (1994) The theory of territory of information: The case of Japanese. *Journal of Pragmatics* 21.1, pp. 67-100.

Kärkkäinen, Elise, 2003. *Epistemic Stance in English Conversations. A Description of its Interactional Functions, with a Focus on I Think*. Amsterdam, Benjamins

Kendon, Adam (2002). Some uses of the head shake. *Gesture* 2.2, pp 147-182.

Kendon, Adam (2004). *Gesture: Visible action as utterance*. Cambridge, Cambridge University Press.

Landmark, Anne Marie Dalby, Pål Gulbrandsen, and Jan Svennevig (2015) Whose decision? Negotiating epistemic and deontic rights in medical treatment decisions. *Journal of Pragmatics* 78: 54-69.

Marín Arrese, Juana (2011). Epistemic legitimizing strategies, commitment and accountability in discourse. *Discourse Studies* 13(6), 789-797.

Marín Arrese, Juana (2015) Epistemicity and Stance: A cross-linguistic study of epistemic stance strategies in journalistic discourse in English and Spanish. *Discourse Studies* 17 (2): 210 –225

McClave, Evelyn Z. (2001). The relationship between spontaneous gestures of the hearing and American Sign Language. *Gesture* 1.1 (2001): 51-72.

Mondada, Lorenza (2013). Displaying, contesting and negotiating epistemic authority in social interaction: descriptions and questions in guided visits. *Discourse Studies* 15 (5): 597-626.

Müller, Cornelia (2004). Forms and uses of the Palm Up Open Hand: A case of a gesture family? In Cornelia Müller and Roland Posner (Eds.), *The Semantics and Pragmatics of everyday Gestures*. Berlin, Weidler.

Poggi, Isabella (2007) *Mind, Hands, Face and Body. A goal and belief view of multimodal communication*. Weidler Buchverlag

Bitti, Pio E. Ricci, Luisa Bonfiglioli, Paolo Melani, Roberto Caterina, Pierluigi Garotti (2014) Expression and communication of doubt/uncertainty through facial expression. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education* 9.1: 159-177.

Robinson, Jeffrey D., and John Heritage (2016). How patients understand physicians' solicitations of additional concerns: implications for up-front agenda setting in primary care. *Health communication* 31 (4): 434-444.

Salager-Meyer, Françoise (1994) Hedges and textual communicative function in medical English written discourse. *English for specific purposes* 13.2: 149-170.

Peräkylä, Anssi (1997) Conversation analysis: a new model of research in doctor-patient communication." *Journal of the Royal society of Medicine* 90.4: 205-221.

Poggi, Isabella, and Francesca D'Errico (2016). Finding Mussolini's charisma in his multimodal discourse. In Fabio Paglieri, Laura Bonelli, and Silvia Felletti (Eds.) *The Psychology of Argument. Cognitive Approaches to Argumentation and Persuasion*. London, College Publications.

Roseano, Paolo, González Montserrat, Joan Borràs-Comes, and Pilar Prieto (2014). Communicating Epistemic Stance: How Speech and Gesture Patterns Reflect Epistemicity and Evidentiality. *Discourse Processes*. DOI:10.1080/0163853X.2014.969137.

Sacks, Harvey (1992) *Lectures on Conversation*. Oxford: Blackwell.

Simon-Vandenbergen, Anne-Marie, and Karin Aijmer. *The semantic field of modal certainty: A corpus-based study of English adverbs*. Vol. 56. Walter de Gruyter, 2007.

Stivers, Tanya, Lorenza Mondada, and Jakob Steensig. "Knowledge, morality and affiliation in social interaction." *The morality of knowledge in conversation*(2011): 3-24.

Streeck, Jürgen (2009). *Gesturecraft: The Manu-facture of Meaning*. Amsterdam, John Benjamins.

Vincze, Laura, and Isabella Poggi (2011). Communicative functions of eye closing behaviours. In Anna Esposito, Alessandro Vinciarelli, Klára Vicsi, Catherine Pelachaud, and Anton Nijholt (Eds.) *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*. Berlin-Heidelberg, Springer. 393-405.

White, Peter (2003) Beyond modality and hedging: A dialogic view of the language of intersubjective stance. Text 23(2, pp. 259–284

Zuczkowski, Andrzej, Ramona Bongelli, and Ilaria Riccioni I (2017). *Epistemic Stance in Dialogue: Knowing, Unknowing, Believing*. Amsterdam, John Benjamins.

Forte, piano, crescendo, diminuendo. Gestures of intensity in orchestra and choir conduction

Isabella Poggi
Dipartimento di Filosofia, Comunicazione e Spettacolo - Università Roma Tre
isabella.poggi@uniroma3.it

Alessandro Ansani
Dipartimento di Filosofia, Comunicazione e Spettacolo – Università Roma Tre
alessandro.ansani@gmail.com

Abstract

Starting from the hypothesis that the signals used in conducting are systematic and shared, so as to make up a lexicon, this work finds out and analyses 23 gesture types, and 11 values in their parameters, that within the body signals of choir and orchestra conductors convey intensity indications: *piano, forte, crescendo, diminuendo*. The gestures selected from a corpus of concerts and rehearsals are analysed through an annotation scheme that describes the body signals soliciting dynamic action, their precise meanings, and the semiotic devices used; for example, enlarging arms progressively, by representing something growing in width, evokes a *crescendo*: an iconic resemblance with transposition across modalities, where amplitude is transposed from the visual to the auditory domain.

1. The conductor's job

The conductor's job is a complex plan of action (Poggi, 2011) aimed at having an ensemble play in a masterly way to convey the enchantment of music. In both concert and rehearsal, the conductor conveys various kinds of information about the sound to produce: who should play or sing, when, what content to express in the words sung, what sound to produce and how. This implies taking care of all the parameters of music: the conductor asks for a particular melody, rhythm, tempo, timbre, intensity, expression, but also reminds of aspects of the piece musical structure, such as coming back to the tonic or turning from minor to major (Ashley 2000; Poggi, 2011). The conductor's work is multifunctional and multimodal: to convey all this information it exploits the whole body, gestures, gaze, head movements, facial expression, posture – and during rehearsals words as well. For all this to be effectively and speedily conveyed during performance, the conductor and the ensemble must share a common language.

The signals used during musical performance are not idiosyncratic but systematic and shared. Concerning a pianist's body movements, a "lexicon" was outlined (Poggi, 2006), i.e., a list of correspondences between signals and meanings. There are communicative signals conveying a performative and a propositional content; expressions of mental states like concentration or cognitive load, and of the emotional states felt about one's playing (e.g., flow or worry) or to be conveyed in music (sadness, mirth). Other body behaviors that typically accompany the technical movement needed to produce a particular sound, timbre, rhythm, intensity, "help" to perform it: e.g. *frowning*, an expression of anger, by mobilizing the energy of this emotion, helps the pianist to play "forte".

Unlike the pianist's body behaviors, the conductor's body signals are all, by definition, communicative, and then more likely to constitute a systematic lexicon. To find out the "lexical items" of such a lexicon, two different paths can be chosen: by modality or by semantic content. In the former case, one tries to find out the meanings conveyed by all signals in a single modality, like did Boyes Braem and Braem (1998) for gestures and Poggi (1998; 2017; in press a) for gaze; in the latter, one may single out a specific class of contents that the conductor must convey during performance, and for each wonder how they are conveyed in whatever modality (Poggi, in press b). This is the approach we adopt in this paper, and the semantic area we investigate in our research concerns the parameter of musical intensity.

2. The multimodal lexicon of musical intensity

This work presents an observational qualitative study aimed to analyse the gestures of intensity, by which orchestra and choir conductors convey the dynamic musical indications "forte", "piano", "crescendo" and "diminuendo". Such indications, that interpret the graphic ones on the author's score, are conveyed to musicians during performance by voice, hands, gaze, face, head, body movements, through requestive acts that can be paraphrased as "sing/play soft" or "loud" (*piano*/*forte*), "progressively lower music/voice intensity" (*diminuendo*) or "progressively make it louder and louder" (*crescendo*).

Eleven fragments of orchestra and choir conduction by three different conductors were analyzed: two, respectively, from a concert and a rehearsal by Riccardo Muti, one from a rehearsal by Leonard Bernstein, and eight (4 from concert and 4 from rehearsal) by Alessandro Anniballi, the conductor of the amateur choir "Orazio Vecchi" of Rome. 130 minutes of fragments in total, 35' of concert and 95' of rehearsal, respectively, were analyzed through the annotation scheme in Table 1.

Here the simultaneous movements in two modalities (posture and gesture) are analysed: col.1 contains the timecode in the video, col. 2 the dynamic indication written on the musical score, 3 (in case of choir performance), the words sung at the same time of the conductor's signal analysed; 4, the modality under analysis; 5, a description of the analysed signal in terms of its parameters; col. 6 contains the goal of the body movement performed, which counts as its "bodily" meaning, i.e., the bodily action from which the communicative meaning stems (see the "originary" meaning in Poggi, 2007): here, *Bust forward, shoulders closed, head forward downward* is the posture of someone bending forward to become smaller. Col. 7 contains the meaning conveyed by the movement described in col.5: *bending forward* to get smaller means: "softer" (i.e., "make a 'smaller' sound"). Col 8. classifies the signal and col.9 clarifies its underlying semiotic device: *making oneself smaller* is an iconic gesture (col.8) that exploits a transmodal shift (col. 9), from space to sound: a body taking less room recalls a sound taking less energy.

Table 1.
An annotation scheme for signals of intensity

1. Time	**2. Score**	**3. Words**	**4. Modality**	**5. Signal description**	**6. Bodily meaning**	**7. Meaning**	**8. Signal type**	**9. Semiotic device**
3.18	*Più piano*	Je-e-su-u Chri-i-**i-ste**	posture	*Bust forward, shoulders closed, head forward downward*	I make myself smaller →	Softer	Iconic	Transmodal iconicity: space→sound Take less room = make softer sound
			gesture	*Elbows folded* *Rh. Open, palm down, pats downward*		please quiet attenuate	Symbolic gesture	Generic Codified

Based on this annotation scheme, with videos in mute mode, a careful analysis was carried out by two independent judges of the conductor's signals of intensity, setting their correspondences with the dynamic indications "forte", "crescendo", "piano", "diminuendo" or other.

In the resulting tentative lexicon, the body signals for intensification are more frequent than ones for attenuation: respectively, 63 for "forte" and 4 for "crescendo", 39 for "piano" and 5 for "diminuendo". Here we focus on intensity gestures, while previous work has overviewed body signals of intensity in all modalities (Poggi, in press).

3. Semiotic devices in intensity gestures

Out of the signals of intensity performed in all modalities, within those performed by hands, arms and shoulders, 21 gesture types were singled out, resulting from the operation of five semiotic mechanisms:

1. **generic symbolic gesture**: the conductor uses a symbolic gesture that could also be understood by non-musicians, being also exploited, with the same meaning, in everyday life by laypeople. E.g., *Index finger over lips*, meaning "be silent", is used to ask for "piano";
2. **specific symbolic gesture**: a symbolic gesture that in everyday life has a certain meaning is used in conduction with a slightly different or a more specific meaning. *Hands, palm up, oscillating on wrist up-down*, that generally means (Morris, 1977; Poggi, 2007) "come here" in conduction means "come on, play/sing louder";
3. **direct iconic gesture**: the conductor's gesture imitates some movement in another modality. E.g., *arms curve widening*, that imitate a swelling body, ask for a "crescendo": a swelling sound. This is a case of "transmodal iconicity" (a sort of synaesthesia), where an analogy is set between widening of a physical shape and progressive amplification of a sound: from a visual domain to an auditory one;
4. **indirect iconic gesture**: the conductor's gesture does not directly imitate the movement it refers to, or its transmodal analogue, but some movement that by inference may recall the desired intensity. Such indirect iconicity may pass through two different kinds of movement:
    a. **motor attitude**: the gesture imitates a movement usually performed while producing another movement or the resulting sound. To mean "sforzato" (forte with effort) the conductor suddenly *clenches his fist*, thus imitating the movement people do when striving in some physical action;
    b. **emotion expression**: the gesture imitates movements typically performed in the expression of an emotion that, when felt, induces the wanted type of attitude or movement. *Hands in claw shape, vibrating with high muscular tension* work as an indication for "forte", because tension is typical of an activating emotion like anger, and anger calls up to the energy required for playing or singing "forte".

In this last case, expression of an emotion (possibly through a body feedback device) evokes the emotion that typically induces the physiological conditions for the right technical movement, thus working as a shortcut to the technical movement. But also, performing the facilitating movement (the emotion expression) becomes a signal requesting that technical movement, i.e., an intensity indication.

4. Meanings of intensity in the global gesture and in its single parameters

Every gesture can be analysed in terms of parameters like handshape, location, orientation and movement (Stokoe, 1978; Volterra, 1987, Calbris, 1990; 2003; Kendon, 1988; 2004; Poggi, 2007), where movement includes, beyond the subparameters of direction, velocity and duration, also the expressivity parameters (Hartman et al., 2002; Poggi, 2007; Poggi and Pelachaud, 2008) of amplitude, tension, fluidity. Each gesture is defined by the values it assumes with respect to all parameters.

Sometimes the dynamic indication is conveyed by the gesture globally, i.e., by the information borne altogether by all of its parameters (GLOBAL GESTURE); but in some cases only one aspect of it bears the intensity indication: e.g., "forte" is not conveyed by the whole gesture but only by the value "*high muscular tension*" within the expressivity parameter of tension (SINGLE PARAMETER).

All in all, the global gesture types for the four intensity indications are 8 for "forte", 7 for "piano", 4 for "crescendo" and 2 for "diminuendo".

In the following, for each indication we first list the global gesture, then its pertinent parameters.

4.1. Gestures requesting to play or sing "forte"

In our corpus, the gesture types for "forte", ordered according to the parameter of handshape, are the following:

1. *Open hand*
    a. *right hand, palm up, oscillates on wrist from musicians to conductor*, as if meaning "come on, come here, come forward"
2. *Extended index finger*:
    a. *both index fingers pushed towards musicians*
3. *Fist* shape:

   a. *right hand, palm to left, pushed forward towards musicians*
   b. *both hands, palms to each other, pushed forward towards musicians*
   c. *both hands, palms down, pushed forward towards musicians* (See Fig.1)
   d. *both hands, palms to each other, pushed downward along hips* (See Fig.2)
   e. *right hand, palm up, moved forward with fluid movement towards musicians*
4. *Claw* shape (*open hand* with *curve fingers* with *high muscular tension*):
   a. *right hand, palm up, moves towards musicians*
   b. *right hand, palm to conductor, vibrating*

In one case, the hand switches from a shape to another:

5. From fist to claw
   a. *right hand in fist shape, palm up, pushed towards musicians opens up in a claw*

That being said, we shall now describe the above gesture types in greater detail:

1.a.: *right hand open, palm up, oscillating on wrist from musicians to conductor*

As mentioned above, some of the conductors' signals are used only in conducting, others also in everyday communication. Gesture 1.a., which means for laypeople encouraging the interlocutor to come closer ("come on, come here"), in conduction may be a request for "forte" for two reasons: first, it conveys encouragement: it means "you are strong, don't be afraid, do dare", hence, finally, "don't be shy, play/sing forte!"; second, since the closer a sound is to your ears, the louder you hear it, requesting "come closer" implies "make your sound be heard louder".

2.a. *both index fingers pushed towards musicians*

This is not a symbolic gesture proper, yet in everyday life it might be interpreted as a peremptory command, a blame or threat of punishment: index fingers violently push forward-downward like daggers or guns, urging the other to do something without discussing the command. The energy impressed to this movement then works as a request for "forte" (Fig.1).

3.a. *right hand in fist shape, palm to left, pushed forward towards musicians*
3.b. *both hands in fist shape, palms to each other, pushed forward towards musicians*
3.c. *both hands in fist shape, palms down, pushed forward towards musicians*

Figure 1
Riccardo Muti: *fists pushed forward* to mean "forte"

3.d. *both hands in fist shape, palms to each other, pushed downward along hips*

Figure 2
Leonard Bernstein: *fists pushed downward* to mean "forte"

These gestures are not generally used in everyday communication, yet in this context their energetic movement conveys a request for high intensity sound, and the variant of both hand vs. one hand corresponds to a gradient of requested intensity.

3.e. *right hand in fist shape, palm up, moved forward with fluid movement towards musicians*

While the gestures above imply a pushing movement with high muscular tension, here the motion is fluid. Yet, the movement towards musicians and the fist shape still ask for higher intensity.

4.a. *right hand in claw shape, palm up, moves towards musicians*
4.b. *right hand in claw shape, palm to conductor, vibrating*

The hand in *claw* shape, i.e., *open hand, fingers curved with high muscular tension*, beside conveying intensity, also indicates a specific way to produce the sound: the muscular tension of the conductor's fingers calls for tense movements by players or tense voice emission by singers.

5.a. *right hand in fist shape, palm up, pushed towards musicians opens up in a claw*

Here the handshape shifts from "*fist*" to "*claw*": the sound must be "forte" but also "tense", possibly vibrating.

Some cases where a single parameter is responsible for the meaning "play/sing forte". Within HANDSHAPE, *"fist"* is prototypically connected with a meaning of strength, since it embeds the visual metaphor (Boyes Braem, 1981) of a firm grasp: strength, power, energy. The *"claw"* shape incorporates the idea of a strong tense grasp. As to DIRECTION OF MOVEMENT, *towards musicians* generally asks for "forte"; so does, sometimes, PALM ORIENTATION, with *palms up* meaning "forte" and *palms down* "piano". In the expressivity parameters, low FLUIDITY and high TENSION typically convey "forte". At times "forte" is expressed by various parameters of the gesture, like in a case from Leonard Bernstein's rehearsal of Stravinskij's "Rite of Spring". To provide a visual image of the dawn of the world, he evokes the dinosaurs, reminding of how heavy their steps could have been, and while conducting, with his arms along his hips, he *pushes both fists downward alternatively* (Fig.2). This gives the image of the dynosaurs' steps, but also scans rhythm and asks for "forte", heavy and tense, by exploiting two parameters: the *fist* handshape, and its *jerky* (low fluidity) *movement*, with its sudden blocked impact.

4.2. Pertinent values in the gesture parameters asking for "forte"

In summary, within the parameters and subparameters of gesture, the values that most typically convey the dynamic indication "forte" are the following:

- within HANDSHAPE, "forte" is most typically conveyed by *fist* and *claw*
- within the subparameters of movement and the expressivity parameters:
    - DIRECTION: "forte" is conveyed by *towards musicians*
    - FLUIDITY: *jerky* as opposed to *fluid* movements
    - TENSION: *high*
    - AMPLITUDE: *wide* movements (for instance, *wide open arms*)
    - QUANTITY OF MOVEMENT: *high*
    - MANNER OF MOVEMENT: *vibrating*

Apparently, these aspects of body movement seem to "naturally" convey ideas of strength and energy.

4.3. Gestures for "piano"

The whole gestures used to request "piano" are the following:

6. Open hand
    a. Both hands open with close fingers, palms down, move inward – outward like smoothing a surface

   b. *Both hands open with close fingers, palms down, slightly move downward, as if keeping a surface down*
   c. *Both hands open with close fingers, palms forward, move forward*
   d. *Both hands open with close fingers, palms forward, oscillate on wrist left-right*
   e. *One or both hands open, palms down, alternatively move fingers up and down gently*
7. Extended index finger:
   a. *Right hand, palm to left, puts extended index finger before mouth*
8. Precision grip:
   a. *right hand or both hands with thumb and index finger touching, palms forward move forward fluidly.*

The *extended index finger touching lips* (n.7) is the symbolic gesture for "keep silent" (Fig. 3).

Figure 3
Alessandro Anniballi: *extended index finger touching lips* to mean "piano"

The *precision grip* (*thumb and index finger touching)* (n.8) carries the visual metaphor of taking something with caution and delicacy. The *open hand* (6), instead, is used in various gestures:

6.a. *Both hands open with close fingers, palms down, move inward – outward* as if smoothing a flat surface

Figure 4
Riccardo Muti: *hands palms down as if smoothing a flat surface* = "piano"

This iconic gesture imitates someone smoothing a flat horizontal surface: a metaphor for something continuous, without any abrupt change, meaning "a sound without any peaks of intensity".

6.b. *Both hands open with close fingers, palms down, slightly move downward,* as if pushing a surface down

One more iconic gesture that means: "keep it down". Again, preventing a surface from coming upward resembles, in a spatial domain, the act of keeping something low in the acoustic domain.

6.c. *Both hands open with close fingers, palms forward, move forward*

As a symbolic gesture, this means "stop there, do not come further"; but since coming closer spatially implies being heard louder, the spatial request not to come closer implies a request for a lower sound.

6.d. *Both hands open with close fingers, palms forward, oscillate on wrist left-right specularly*

This symbolic gesture is an augmentative form of *shaking index finger* to mean "No" (not only one finger, but both whole hands shaking): an emphatic negative request, "don't do this", which here means "don't sing/play so loud".

6.e. *One or both hands open, palms down, alternatively move fingers up and down gently*

This gesture is not used in laypeople's everyday communication. The shape is the open hand, but fingers move as if gently playing an imaginary piano, conveying an image of a surface that moves in a gentle and non-conspicuous way, like the twinkling of a calm sea under the moon.

4.4. Pertinent values in the parameters of the gestures for "piano"

Within the expressivity parameters of movement, those calling for a decrease in musical intensity are the following:

- FLUIDITY: movements are always very *fluid*: hands are often rocking, almost dancing
- TENSION: *low*
- AMPLITUDE: *low*
- QUANTITY OF MOVEMENT: *low*

All these values are related to a decrease in energy, which points at decrease in sound intensity. Lower energy implies lower quantity of movement, lower amplitude and tension, and higher fluidity.
Interestingly enough, the gestures for "forte" and "piano" are characterized by opposite values in their parameters: "forte" by fist handshape, high muscular tension, jerky movement, high quantity of movement, upward direction, while "piano" by flat hand, low tension and high fluidity, low quantity of movement, downward direction. This reminds of Darwin's (1872) law of opposition, according to which opposite movements correspond to opposite emotions (or opposite meanings?).

4.5. Gestures for "crescendo" and "diminuendo"

Out of the four gestures for "crescendo" in our corpus, in three of them the hand is *open* and *loose*, one is in the *claw* shape.

9. Open hand curve loose
   a. Arms with open hands curve open outward while shoulders raise upward

Figure 5
Alessandro Anniballi: swelling movement = "crescendo"

   b. Left hand open curve loose, palm to conductor, rotates forward repeatedly in wider and wider rounds
   c. *Both arms open curve loose, palm up, alternatively move fingers up and down gently towards conductor*

Figure 6
Alessandro Anniballi: *fingers gently moving up and down* = "crescendo"

Gestures 9.a. and 9.b. exploit the visual metaphor of something swelling; 9.c., like gesture 6.e above, imitates something moving and changing softly.

10. Claw shape
    a. *right hand in claw shape, palm down, raises upward*

Here the high tension of the claw handshape parallels the tension of progressively increasing intensity. The two gestures for "diminuendo" are the following:

11. Open hand loose
    a. *left arm with open hand palm down retracts backward*

12. V shape (*index and middle fingers extended open*)
    a. *right hand in V shape, palm to conductor, moves rightward progressively closing index and middle finger*

The first gesture, mainly in its parameters of movement (retraction) and direction (backward) conveys the idea that something previously larger should now decrease. The second is iconic too, but one that does not imitate the musicians' movements or the perception of their acoustic effects; it imitates the graphic symbol used in musical scores for the dynamic indication of "diminuendo": >, called "forchetta" (fork), that graphically represents something which first is larger and then becomes thinner; with "forchetta" being in its turn an imitation – through graphic means – of lowering acoustic intensity.

5. Conclusions

Our study aimed at outlining the lexicon of intensity in the gestural language of orchestra and choir conductors. The recurrence of the same gestures, or of the same values in their parameters, in different conductors, shows that intensity gestures constitute a specific lexical area within the system of musical indications, governed by recurrent and systematic semiotic devices.

These gestures do not make up a specialist jargon, rather they are quite similar to everyday gestures and likely comprehensible by laypeople. Among the symbolic gestures used, only a few of them have a more specific meaning, and the direct and indirect iconic gestures exploit the same mechanisms for gestural creation as plain language, such as metaphor and metonymy. When a conductor widens his arms imitating a swelling body, he is transferring some properties of something in the visual domain onto something in the acoustic domain: by communicating: "Swell the sound the same way I swell my body", he relies on the mechanism of metaphor. On the other side, when he clenches his fists he implicitly conveys "sing/play with the same strength I express with the muscular tension of my fists", here exploiting a metonymy.

Those that we call rhetorical figures are in fact powerful devices for the creation of new signals, and their operation can be found in many of the conductor's gestures. This again reinforces our hypothesis of continuity between gestures for musical indications and everyday gestures.

The study of bodily signals in musical performance will be widened and deepened in future. First, the signals for intensity in other modalities, e.g. gaze, will be investigated. Second, other musical indications – like expressiveness, timbre or rhythm – will be overviewed, and the possible polysemy of some signals will be studied in depth; in fact, sometimes the same gesture or gaze may be ambiguous between conveying intensity or simply rhythmic accent. Finally, moving from observational to empirical studies, our hypotheses on the meanings of single gestures and other signals will be tested by asking laypeople and musicians to guess the meanings of conductors' gestures: this will tell us if the supposed iconicity of some gestures may be of help to comprehension even to non-expert people, and if the lexicon of conduction is not so tightly codified as to be regarded as an overspecialized language.

Acknowledgments. We are indebted to Mº Alessandro Anniballi for his inspiring gestures and his generous acceptance for them to become an object of study.

References

Ashley, R. (2000). The pragmatics of conducting: Analyzing and interpreting conductors' expressive gestures. In C. Woods, G. Luck, R. Brochard, F. Seddon and J. A. Sloboda (Eds.), *International Conference on Music Perception and Cognition, Keele, UK,* (*2000*).

Boyes Braem, P. (1981). *Significant Features of the handshape in American Sign Language*. Unpublished PhD. Thesis, University of California, Berkeley.

Boyes Braem, P. and Th. Braem (2004). Expressive gestures used by classical orchestra conductors. In C. Müller and R.Posner. (Eds.), *The Semantics and Pragmatics of Everyday Gestures*. Berlin: Weidler.

Calbris, G. (1990). *The semiotics of French Gestures*. Bloomington, Indiana: Indiana University Press.

Calbris, G. (2003). *L'expression gestuelle de la pensée d'un homme politique*. Paris: Ed. du CNRS.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. New York and London: Appleton and Company.

Hartmann, B., M.Mancini and C. Pelachaud (2002). Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis. *Computer Animation 2002*, 111-119.

Kendon, A. (1988). *Sign Languages of Aboriginal Australia. Cultural, semiotic, and communicative perspectives*. Cambridge: Cambridge University Press.

Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge: Cambridge University Press.

Poggi, I. (2006). Body and mind in the Pianist's performance. In M.Baroni, M.R.Addessi, R.Caterina and M.Costa (Eds.), *Proceedings of the 9th International Conference on Music Perception and Cognition*, ICMPC, Bologna, 22-26 August, 2006, pp.1044-1051.

Poggi, I. (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication*. Berlin: Weidler.

Poggi, I. (2011). Music and leadership: the Choir Conductor's multimodal communication. In M.Ichino and G.Stam (Eds.), *Integrating Gestures. The interdisciplinary nature of gestures*. Amsterdam: John Benjamins, 2011, pp.341-353.

Poggi, I. (2017). Gaze in music performance. Conductors' styles of eye communication during concert and rehearsal. Poster at the International Conference *"Language as a form of Action"*, Roma, 21-23 June, 2017.

Poggi, I. (in press a). Lo sguardo del Maestro. I segnali comunicativi degli occhi nella direzione di coro e d'orchestra. In C.Corradi (Ed.) *Scritti in onore di Roberto Cipriani.* Roma: Morlacchi.

Poggi, I. (in press b). Signals of intensification and attenuation in orchestra and choir conduction. *Normas.*

Poggi, I. and C. Pelachaud (2008). Persuasion and the expressivity of gestures in humans and machines. In I.Wachsmuth, M.Lenzen and G.Knoblich (Eds.), *Embodied Communication in Humans and Machines*. Oxford, Oxford University Press, p.391-424.

Stokoe, W.C. (1978). *Sign Language Structure: An Outline of the Communicative Systems of the American Deaf*. Silver Spring: Linstock Press.

Volterra, V. (1987). *LIS. La Lingua Italiana dei Segni*. Bologna: Il Mulino.

The human body in multimodal communication: the semiotic conceptualization of hair

Grigory Kreydlin
Russian State University for the Humanities
Moscow, Russia
gekr@iitp.ru

Lidia Khesed
Russian State University for the Humanities
Moscow, Russia
lidakhe@yandex.ru

The article presents a fragment of the semiotic conceptualization of the somatic object "hair <on the head>" in the Russian language and nonverbal code. We describe the meaning of the standard Russian nomination of hair – *volosy*, identify and analyze key features of hair (such as its size, shape, colour, structure, functions and dysfunctions), the meaning of these features and their standard Russian nominations. A separate part of the article is devoted to the gestures performed with hair, namely caressing and aggressive gestures. By the comparative analysis of verbal and nonverbal Russian sign units we explore connections between characteristics of hair and the corresponding communicative behavior of the person. Special attention is paid to the features of hair which are either semantically or culturally marked as their meanings reflect either social or cultural stereotypes peculiar to Russians. Observation of some Russian set phrases and idioms with the word *volosy* reveals the problem of literary translation of the corresponding texts into foreign languages.

Index terms: body, gestures, hair, Russian, semiotic conceptualization

The present article[1] continues a series of the research devoted to the construction of the **semiotic conceptualization of the human body**. This concept and the term reflect how the **corporal, or somatic, objects** are represented in a natural language (Russian in particular) and how they are used in the corresponding body language. The latter includes several types of nonverbal sign units, such as manual gestures, head and shoulder gestures, postures, facial expressions, etc. [6, 13]. The semiotic conceptualization of somatic objects and corporeality is a formal model that describes what ordinary people, non-specialists, think and say about different somatic objects and how they use them in performing gestures in various patterns of corporeal communicative behaviour.

Somatic objects can be subdivided into several classes. These are the body itself and its parts (e.g. head and arm), parts of body parts (fingers, nostrils) parts of parts of body parts and so on. In

[1] The research is conducted within the framework of two research projects:
Corporeal Manifestations of Mental and Emotional Activity of the Human Being project supported by grant No. 16-34-00023 of the Russian Academic Foundation for the Humanities and Russian Somatic Vocabulary: Cognitive and Semiotic Aspects project supported by grant 16-04-00051 of the Russian Academic Foundation for the Humanities.

Russian, the depth of subdivision is 6. The rest are organs (e.g. liver or heart), corporeal liquids (blood, sweat, tears), special places on the human body (armpit, navel), lines (waist, treats) etc.

Semiotic conceptualization of the human body combines different constituents that form several classes. These are (a) somatic objects and (b) their names, (c) features of the somatic objects, (d) values of the features and (e) their names, (f) gestures produced by somatic objects and some other constituents [8: 230 – 234; 9: 42 – 51]. The description of a somatic object is a fragment of the semiotic conceptualization of the human body.

The aim of our article is to describe the object that is called 'hair <on the head>' or *vólosy* in Russian. This infers displaying the meaning of the word, its synonyms and derivatives and also to consequently present its physical, structural and functional features, underlining the role that hair plays in Russian culture and aesthetics.

Through this type of description, we explore several aspects of multimodal communication. It connects units of different systems and codes together. Hair serves as an object of both verbal and nonverbal interaction. The units of both codes shape a number of practices involving hair, such as care, treatment and rituals with hair. Caressing or aggressive gestures that involve hair are quite often accompanied by particular expressions, marking the attitude of the person performing the gesture to the addressee. Also, hair as a somatic object connects several systems of the human being (in terminology of [1]). Being a corporeal cover, it is a part of the physical, or bodily, system. At the same time, gestures involving hair and Russian expressions with the word *vólosy* can say a lot about the person's mood, nature of their emotional state, e.g. *volosy vstali na golove <ot straha>* 'the hair stands on end <because of fear>', *rvat' na sebe volosy <v otchayanii>* 'to pull out one's hair <in disdain>'. In these instances, hair is made to be part of the psychic system of the human being. Correlation with the mental system is reflected in a number of thoughtful gestures with hair, such as **to thoughtfully rumple one's hair** or **to thoughtfully bite one's hair lock**. Hair belongs to the protective system as well because of its functions. One of these functions is to protect the head from external influences. In the present research, we define and describe these correlations.

While constructing the semiotic conceptualization of hair, we use the so-called **feature based approach** [7, 12]. The idea and main contents of this approach are to present the somatic object as a set of different features together with different language names of the object itself.

Hair alongside skin and nails forms the class of *body covers*. It is one of the main objects that have a social function of presenting a human being, human appearance and identity. Many features of hair are **semantically** and **culturally marked**, or **salient** [7].

The **semantically marked value** of the feature is the value that characterizes not only the somatic object itself, but the possessor of the object. For example, the value *krivýe* 'ham-handed' of the Russian word *rúki* 'hands' is semantically marked. Thus, the sentence *U Péti krivýe rúki* 'Petya is ham-handed' describes the shape of Petya's hands and presents Petya as awkward and clumsy.

The **culturally marked value** of the feature is the value that both characterizes the somatic object and also reflects some aspects of the culture. For example, the feature "colour of cheeks" has two values: one is /pink/, another is /white/. Both features are marked in Russian culture, because they express not only the colour of the cheeks, but also display the typical Russian stereotypes of human health and illness respectively.

The Russian word *vólosy* 'hair' has several meanings. The first one, or the lexeme VOLOSY 1, can be described as 'thread-like thin somatic object, attached to the head with one end'. The lexeme VOLOSY 1 is a plural form of the lexeme VOLOS 1 but the latter is used more rarely. A typical usage of the lexeme VOLOSY 1 can be illustrated with the sentence (1) *Yá nashlá u sebyá pyát sedýh volós* 'I found five grey hairs on my head'[2]. This context shows that the word *vólosy* in the VOLOSY 1 meaning is countable.

The second meaning of the same word is 'a multitude of VOLOSY 1' (lexeme VOLOSY 2). Here *vólosy* is Pl. Tant. The word *vólosy* in theVOLOSY 2 meaning belongs to **mass nouns**, because it denotes an indivisible and innumerable set of objects. Not all Russian explanatory dictionaries [4, 10, 11, 15] fix this meaning, with the pleasant exclusion being the old four-volume dictionary of the language of Pushkin [18]. Phrases such as *gustýe vólosy* 'thick hair', *uhód za volosámi* 'hair care' or *secrét roskóshnyh volós* 'secret of the marvelous hair' illustrate the usage of the lexeme.

When speaking about hair, people usually mention its size or shape. Size, shape and colour form a triad that embraces the aesthetic perception of a person's appearance. It determines an individual's tidiness or untidiness, attractiveness or shabbiness, beauty or ugliness.

The size of hair can be either **absolute** or **relative**. The absolute size of the somatic object implies its size regardless of any spatial axis, whereas relative size refers to its size in relation to one of the three spatial axes.

The absolute size of hair in Russian is the size of *vólosy* in the VOLOSY 1 meaning ('long thin threads…'). The absolute size is predetermined by the position of the hair on the head, which goes from the top down. In other words, the absolute size of hair is its length, and Russians say *dlínnye vólosy* 'long hair' or *korótkie vólosy* 'short hair'.

The absolute size of hair in the VOLOSY 1 meaning is closely connected with some other features. First, it is the quantity of hair in the VOLOSY 2 meaning. The typical Russian expressions describing the quantity of hair are *mnógo volós* 'large quantity' or *málo volós* 'small quantity'. The second and third are the density and the volume of hair.

[2] All the examples are taken from The National Corpus of Russian Language [14], translation is ours (– authors).

Density expresses how closely the separate thin threads of hair are located to one another and how many of them lie on the head. The top of the density scale is expressed in Russian by the adjective *gustóy* 'thick' and the opposite point of the scale is expressed by the adjectives *rédkiy* 'thin' or *zhídkiy* 'scanty', for example (2) *Sn'áv shápku, jerósha rúsye gustýe vólosy, komissár stál diktovát'* (A. Bek. Pózdnij chas) 'The commissar took off the hat and, rumpling his thick blonde hair, began to dictate' (A. Bek. The Late Hour).

The volume of hair refers to how much space the hair occupies. This feature also contains two opposite values: /large volume/ and /small volume/. Russians have a special word to express a large volume of hair. It is the adjective *pýshnyj* 'voluminous', cf. *pýshnye volosy* 'voluminous hair'. There is no idiomatic mode of expressions for /small volume/. Voluminous hair is perceived in Russian culture as beautiful. It is not surprising that there is a special practice to increase the volume of hair and a special name for this practice – *pridát' ob'jóm volosám*. By increasing the volume of hair, people make hair more attractive and raise its aesthetic appeal.

The relative size of hair is its length and breadth. The hair's length is expressed with the adjectives *dlínnyje* 'long' and *korótkije* 'short', but also with some more complex constructions, cf. (3) *Devítsa v <…> pilótke na volosáh, pohózhih na mélkie zhóltye strúzhki* (Y. Buyda. Shchína). 'A girl wearing a <…> garrison cap on the yellow scob-like hair' (Y. Buyda. Shchina).

The expression *dlínnyje vólosy* 'long hair' does not imply that the hair's length dominates its width. It refers to the hair of the standing person with the hair going top down and describes the hair that is longer than normal hair should be. The expressions *dlínnyje rúki* ('long arms') and *dlínnyje* nógi ('long legs') have the same scheme of explication regarding the meaning.

The relative size of the somatic objects that can't be marked as *dlínnyj* 'long' or *korótkij* 'short' do not have idiomatic expressions. In the National Corpus of Russian Language (NCRL) there are only five entrances of the expression *vólosy srédnej dliný* 'mid-length hair', which also denotes its relative (normal) size. But this is not an idiomatic mode of expression.

The breadth of hair corresponds to the horizontal axis. Although hair can be understood as a number of thin threads, we can describe its particular breadth, cf. *tólstyje vólosy* 'thick hair' and *tónkije vólosy* 'thin hair'. There are also special expressions that describe the increase or decrease of the breadth of the hair, such as *uvelíchit' tolshchinú volós* 'to increase the breadth of the hair' and *umén'shit' tolshchinú volós* 'to decrease the breadth of the hair'.

During the human life the hair grows, and the ability to grow is one of its fundamental characteristics. It is not an accident that some Russian dictionaries mention this property in the explication of the meaning of the word *volosy*, though it makes the lexicographic description excessive.

The shape can be either natural or artificial, which a person creates through different manipulations with the hair.

One of the kinds of shape that can be either natural or artificial is curly hair. In Russian, there are several expressions for 'curly hair', such as *kudr'ávyje vólosy*, *kurchávyje vólosy* or *kucher'ávyje vólosy*. Though these adjectives are synonyms, they have different derivational potential. For example, there is a noun *kúdri* 'curls', which means a separate part of the shape. Two other adjectives do not possess this quality.

Another Russian expression for the shape of hair is *pryamýje vólosy* 'straight hair'. The straight hair falls, hangs down or cascades parallel to the torso, cf. (4) *Pryamýje zhóstkije vólosy pádali u negó <…> na bróvi* (I. Kuprin Konokrády) 'The straight coarse hair fell <…> on his eyebrows' (I. Kuprin. Horse-stealers).

Natural and artificial shapes of hair can change because of different factors – either external (weather, environment, etc.) or internal (age, illness, emotional state of the person, etc.), e.g. (5) *Kózha shchók nalilás róvnym rózovym tsvétom, lób stál bél i chíst, a parikmáherskaja zavívka volós razvilás'* (M. Bulgakov. Máster i Margaríta) 'The skin of her cheeks was evenly suffused with pink, her brow had become white and smooth and the frizzy, artificial wave in her hair had straightened out' (M. Bulgakov. The Master and Margarita. Tr. by M. Glenny. 1967) and (6) *U reb'ónka zakudr'ávilis' vólosy* 'Kid's hair has curled up'.

Colour and tone are also physical features of hair [5]. In most of the European languages, different colours of hair have special nominations. Russian is not an exception here, cf. *kashtánovyje / rýzhyje / rúsyje vólosy* 'brown / red / blond hair' where the adjectives express different meanings of the feature "colour of the hair". The constructions *svétlyje / t'ómnyje vólosy* 'fair / dark hair' and *bl'óklyje / túsklyje vólosy* 'flat / dull hair' denote different tones of the hair. Colours and tones create the basis for the well-known classification of people, cf. such names as *blondín* 'blond', *belokúryj* 'fair haired' for a person with blonde hair, *shatén* for a person with brown hair and *rýzhyj / ryzhevolósyj* for a person with red hair.

Some colours, such as red hair (Rus. *rýzhyje vólosy*), are marked in Russian culture. People with red hair are often perceived with suspicion, as cunning and sly, for example *rýzhyj-besstýzhyj* lit. 'blushless-red' or *Chto ya rýzhyj, chto li* lit, 'Am I red, really?' etc. Grey hair (Rus. *sedýje vólosy*) is also marked. Grey hair is peculiar to old people, but it cannot only mark the age of the person. It is associated with human wisdom, rich life experience and some other features. Thus, the frequent Russian expression *dozhít' do sedýh volós* lit. 'live to a grey hair' reflects not only the colour of the hair but also the positive attitude towards its possessor.

Hair, regarded as a thread (lexeme VOLOSY 1), has two salient parts, which have idiomatic names in Russian – *kóren'* 'root' and *kónchik* 'end'. Hair as a set of elements (lexeme VOLOSY 2) also has parts, but their nominations are different from the previous ones. These are *pr'ád' <volós>* 'hair lock', *puchók <volós>* 'tuft of hair', *vihór* 'forelock', *chúb* 'scalp lock', *klók* 'clump', etc. These

words form a synonymic row with the common component ‘relatively small set, or part, of hair’, but their meanings are different.

Pr’ád’ <*volós*> ‘hair lock’ is ‘a relatively small part of thin threads on the head, which are adjoined to each other and thus perceived as a whole’. *Puchók* <*volós*> ‘tuft of hair’ is ‘a relatively small part of thin threads on the head, which are adjoined to each other on one end and stick up on another end’.

Examples:

(9) *Podstrízhennaja pr’ád’ volós, spúshchennaja na lób, ozhachájet sujetnúju mélochnost’* (A. Chekhov. Rukovódstvo dl’a zhelájushchih zhenítsa) ‘A cut hair lock falling on the forehead marks earthly pettiness‘ (A. Chekhov. A guide for those who want to get married);

(10) *Vólosy torchát neróvnymi puchkámi* (M. Petrosyán. Dóm, v kotórom…) ‘Hair sticks up with rough tufts’ (M. Petrosyán. A house, in which…).

Klók ‘clump’ means ‘a relatively small part of thin threads, which are perceived as a single whole and which are either adjoined to the head with one end or not’.

These explications improve lexicographic descriptions of these words given in Russian dictionaries, because they display the significant semantic characteristics of those words which are absent there. Furthermore, they show that *pr’ád’* and *puchók* are semantically closer to each other than each of them to the word *klók*. Both *pr’ád’* ‘hair lock’ and *puchók* ‘tuft of hair’ reflect the specific location of the part of the hair and its structure, either the inner one or outer one.

Functional features of any somatic object are of two kinds: proper functions and dysfunctions [2: 41 – 54, 9: 42 – 51].

Three main functions of hair are **aesthetic**, **masking** and **protective**.

Aesthetic estimation of the human appearance is based much on the perception of hair. If the hair is dirty, untidy, messy or ugly, the perception of the person is estimated negatively. That is why people care about their hair. Hair decoration, a haircut, the changing of its colour or adding volume are special operations which are performed with hair in order to make the person more pleasant and handsome.

The masking function of hair manifests itself in its ability to hide or mask anomalies of the head, ears, neck and some other somatic objects. Hair may serve to cover unpleasant ears, scars on the forehead, wrinkles, etc. For example, (11) *Strízhka, iméjushchaja pýshnyje lókony – éto optimál’nyj variánt dl’a úzkih líts* ‘A haircut with big curls is the best solution for a thin face’ (forum www.raykovstudio.ru).

As to the protective function of hair, it protects the head and its parts from different mechanical and environmental damages, cf. (12) <*Zimój*> *vólosy sámi sozdajút normál’nuju dl’a*

golový temperatúru i golová ne m'órznet 'In winter the hair keeps the warm temperature of the head, that's why it does not feel cold'(forum www.lovehate.ru).

Some dysfunctions of hair manifest the inner illness of the human body or that of its parts. It is not an accident that hair illnesses constitute a special medical discipline called trichology. Hair illnesses presuppose the special methods of treatment, which usually take a lot of time. If the hair treatment has not been started in time, the illness can transfer to other somatic objects, which are spatially connected with hair. Some illnesses cause hair loss, and some lead to its excessive pathological growth. There are also several kinds of illnesses, the main symptom of which is the appearance of contagious objects in the hair, for example dandruff or lice.

Hair illnesses have both scientific (medical, biological, etc.) and common names. Thus, the common Russian name for the hair illness *seboréja* 'seborrhoea' is *pérkhot'* 'dandruff', and the common Russian name for *gipertrihóz* 'trichauxis' is *volosátost'* 'hairiness'.

Gesture somatism is a nonverbal sign unit, of which the nomination includes the name of the somatic object, the name of its feature or the name of a value of the feature. Gesture somatisms with hair are divided into two groups: one with hair as an active object of the gesture being performed and another with hair as a passive object [7: 116].

Russian cultural norms present hair as non-controlled by people [3, 16, 17]. It grows, falls out, moves, stands on end regardless of a person's will, but in the majority of corporeal sign movements hair is a passive somatic object. Among Russian gestures with hair the most frequent are **pogládit' po volosám / po golové** 'to stroke one's hair / head'; **trógat' vólosy** 'to touch one's hair'; **igrát' chjími-to volosámi** 'to play with one's hair'; **utknútsa litsóm v vólosy** 'to burrow one's face into one's hair'; **tselovát' vólosy** 'to kiss one's hair' and **namátyvat' vólosy / pr'ád' volós na pálets** 'to wind a hair lock around one's finger'. The physical representation of all these gestures includes the act of hair touching. In Russian culture, it is not severely tabooed – that explains why there are lots of gestures with hair touching.

In other cultures, this action is considered as taboo and special permission from the addressee to touch his / her hair is always required, even in a silent form. The permission to touch the hair demonstrates the high degree of trust to the gesture performer. It especially concerns the attitude of a man towards a woman and vice versa, where hair touching expresses feelings of tenderness, love and intimate relations between the two people.

Russian caressing gestures with hair are opposed to Russian aggressive gestures. The goal of aggressive gestures is to bring physical damage to the addressee and to hurt him / her. By doing this, the performer demonstrates an extremely negative attitude to the addressee. The examples of such gestures are **taskát' zá volosy** 'to drag by the hair'; **drat' zá volosy** 'to pull by the hair'; **d'órgat' zá volosy** 'to pull one's hair'; **hvatát' zá volosy** 'to seize one's hair'.

We presented a fragment of the semiotic conceptualization of hair as it is expressed in the Russian language and body language. We have analyzed Russian nominations of hair, considered the physical, structural and functional features of hair and their values. We discussed several types of units related to hair, such as phraseological expressions with the word *vólosy* and gesture somatisms with hair.

Many aspects of the semiotic conceptualization of hair have been left behind. This is the description of symbolic meanings and typology of various corporeal practices with hair. It is also the comparative analysis of different semiotic conceptualizations of hair that are generated by different types of discourse and texts (scientific, folklore, artistic, etc.) as well as the comparative analysis of semiotic conceptualizations of hair in different languages and body languages (e.g. English, French and German). All of these important and interesting problems deserve separate studies[3].

References

1. Apresyan 1995 – Apresyan,Y.D. Obraz cheloveka po dannym yazyka: popytka sistemnogo opisaniya [The Image of a Human Being in the Language: an Essay of Systematic Description] // Izbrannye trudy [Selected works]. Vol. II. Integral'noe opisanie jazyka i sistemnaya leksikografiya [Integral Description of Language and System Lexicography]. – Moscow: Yazyki russkoy kul'tury. P. 348 – 388.
2. Arkadyev, Kreydlin 2011 – Arkadyev, P.M., Kreydlin, G.E. Chasti tela i ih funktsii [Body parts and their functions]. Slovo i yazyk: sbornik k vos'midesyatiletiyu academika Y.D. Apresyana [Collections of works devoted to the 80th anniversary of Y.D. Apresyan]. – Moscow: Yazyki slavyanskih kul'tur. P. 41 – 54.
3. Baranov, Dobrovol'skiy 2007 – Baranov, A.N., Dobrovol'skiy, D.O. Slovar' – thesaurus sovremennoy russkoy idiomatiki [Thesaurus dictionary for Russian idioms]. – Moscow: Avanta+. 1135 p.
4. Dal' 1994 – Dal', V.I. Tolkovyy slovar' zhivogo velikorusskogo yazyka [Russian explanatory dictionary], 4 volumes. V. 1. – Moscow: Terra. 784 p.
5. Kadykova, Kreydlin 2010 – Kadykova, A.G., Kreydlin, G.E. Chasti tela v russkom yazyke i v russkoy kul'ture: priznak "tsvet" [Body parts in Russian language and Russian culture: "colour" feature], Vestnik RGGU (Moskovskiy lingvisticheskiy zhurnal) [Herald RSUH (Moscow linguistic journal)], no.9/52. P 47 – 64.
6. Kendon 1990 – Kendon, A. Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge University Press. 1990. 292 – P.; Kendon 2004 – Kendon A. Gesture. Visible Action as Utterance. Cambridge University Press. 400 p.

[3] The authors express gratitude to prof. Irina Yevseeva and Svetlana Pereverzeva for their attentive and friendly discussing of the paper and for valuable comments.

7. Kreydlin 2002 – Kreydlin, G.E. Neverbal'naya semiotica. Yazyk tela i estestvennyj yazyk [Nonverbal semiotics. Body language and natural language]. – Moscow: NLO. 581 p.
8. Kreydlin 2010 – Kreydlin, G.E. Telo v dialoge: semioyicheskaya kontseptualizatsiya tela (itogi proekta). Chast' 1: telo i drugie somaticheskie ob''ekty [Body in the Dialog: Semiotic Conceptualization of the Body (Results of the Project). Part 1: Body and Other Somatic Objects]. Komp'yuternaya lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnogo seminara "Dialog" [Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference "Dialog"] (Bekasovo, 26–30 May 2010). Vol. 9 (16). – Moscow: RSUH publishing house. P. 230 – 234.
9. Kreydlin, Pereverzeva 2010 – Kreydlin, G.E., Pereverzeva, S.I. Semioticheskaya kontseptualizatsiya tela i yego chastey. Klassifikatsionnye i structurnye harakteristiki somaticheskih objektov [Semitotic conceptualization of the human body and its parts. Classificatory and structural features of the somatic objects]. Voprosy filologii [questions of philology]. № 2 (35). P. 42 – 51.
10. Kuznetsov 1998 – Kuznetsov, S.A. Bol'shoy tolkovyy slovar' russkogo yazyka [Big Russian explanatory dictionary]. – Saint-Petersburg: Norint. 1536 p.
11. Evgenieva 1999 – Evgenieva, A.P. Slovar' russkogo yazyka [Dictionary of Russian language].Vol. 1–4. URL: http://www.slovari.ru/default.aspx?p=240.
12. Müller, Cienki 2008 – Metaphor and Gesture. Cienki, A., Müller, C. eds. Amsterdam: John Benjamins. 307 p.
13. Poggi 2001 – Poggi, I. Towards the Alphabet and the Lexicon of Gesture, Gaze and Touch. In: Multimodality of Human Communication. Theories, problems and applications. Virtual Symposium edited by P. Bouissac. URL: http://www.semioticon.com/virtuals/index.html.
14. Ruscorpora 2004 – 2017 – National Corpus of Russian Language. URL: http://www.ruscorpora.ru/.
15. Ozhegov, Shvedova 1992 – Ozhegov, S.I., Shvedova, N.Y. Tolkovyy slovar' russkogo yazyka [Russian explanatory dictionary]. – Moscow, 1992. URL: http://www.ozhegov.org/.
16. Telia 2006 – Telia, V.N. Bol'shoy frazeologicheskiy slovar' russkogo yazyka [Big Russian phraseological dictionary]. – Moscow: AST. 784 p.
17. Tolstaya 2008 – Tolstaya, S.M. Prostranstvo slova. Leksicheskaia semantika v obshcheslavianskoi perspective [Word space. Lexical semantics in the Slavic languages]. – Moscow: Indrik Publ. 528 p.
18. Vinogradov 2000 – Vinogradov, V.V. Slovar' yazyka Pushkina [Dictionary of Pushkin's language], 4 volumes. V.1. – Moscow: Azbukovnik. 976 p.