The 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen 29-30 September 2016: Extended Abstracts

Merel Jung, Mannes Poel and Dirk Heylen. *Living with a robot pet: context-based interpretation of social touch in human-robot interaction*

Lidia Khesed and Grigory Kreydlin. *The human body in multimodal communication: semiotic conceptualization of hair*

Varduhi Nanyan. *Does L2 speech generate a higher gesture rate? A study of Dutch speakers of English*

Costanza Navarretta. *Speech pauses, gestures and audience laughter in English humorous speech*

Thomas Ousterhout. *BCI effectiveness test through N400 replication study*

Thomas Ousterhout. *Investigation of the semantic priming effect with the N400 using symbolic pictures in text*

Matthias A. Priesters, Kirsten Bergmann and Stefan Kopp. *Deriving Individual Gesture Profiles from a Multimodal Dialogue Corpus*

Isabella Poggi. *Forte, piano, crescendo, diminuendo: signals of intensification and attenuation in orchestra and choir conduction*

Kalin Stefanov and Jonas Beskow. *Gesture Recognition System for Isolated Signtem for Isolated Sign*

Ye Tian, Chiara Mazzocconi and Jonathan Ginzburg. *Time alignment between laughter and laughable*

Laura Vincze and Isabella Poggi. *A repertoire of multimodal signals communicating a speaker's overall epistemic stance*

Bjørn Wessel-Tolvig and Patrizia Paggio. *Can co-speech gesture change the perception of ambiguous motion events? Experimental evidence from Italian*

Semiotic and Modality Diversity in Languaging: Implications for Some Debates about Language Origins

Adam Kendon

University of Cambridge and University College London

In language origins discussions the debate between "gesture firsters" and "speech firsters" has arisen because all involved think of "language" as an abstract semiotic system, monomodalic in its realisation: *either* kinesic (sign/gestural) or oral-aural (speech). The "gesture firsters" think that humans must have "switched" from sign language to spoken language, but cannot well explain how or why this happened. The "speech firsters", while they have no "switch" problem, largely ignore gesture use and have yet to integrate sign languages into any language origins scenario. However, close observation of a person engaged in *languaging* shows that there is always at play an orchestrated *diversity* of articulations and semiotic processes (languaging is multimodal!). This must always have been so. Over the eons during which capacities and methods of symbolic or referential communication elaborated, a differentiation into partially separated specialised systems took place. "Language" as conceived of since Saussure, for example, is one such specialisation. The question of its origin can be re-cast as a question about the processes of differentiation and specialisation which, it will be seen, are as much the outcome of social processes as they are of processes of biology.

Multimodal health communication in two cultures – A comparison of Swedish and Malaysian Youtube videos

Jens Allwood, Elisabeth Ahlsén, Stefano Lanzini, Ali Attaran
University of Gothenburg & University of Malaya, Kuala Lumpur

Introduction: The paper presents a comparative study of multimodal communication in Youtube videos on obesity and health in Sweden and Malaysia, containing information and propaganda about obesity. The main message is to make people want to lose weight and to inform about and/or sell ways of doing this. Since there are major cultural differences between Sweden and Malaysia, this 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.

Method: 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 and analyzed further, using rhetorical analysis.

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, Bean, & Johnson, 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). This appeal to attitudes and emotions can further be approached from two perspectives: *expressive* and *evocative* (Allwood, 1978).

The Swedish videos were 1) "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) "Fighting against obesity" - a commercial for clinic with a specific operation technique for obesity, 3) Weight Line 1 - an information video about health and obesity from a health website with a famous doctor, and 4) Weight Line 2 - a video from a health website, with a famous doctor promoting a specific diet.
The Malaysian Youtube videos were: 1) a short animated film (the bursting balloon) – creator unknown, 2) a information film from the Health Ministry of Malaysia about obesity, 3) ”The fat nation" – a commercial from a well-known gym chain, 4) ”Obesity in Malaysia", promoting slimming product, and 5) "Malaysian obesity" a propaganda film warning about obesity – creator unknown.

Results: ***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. One video lacks logos.

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 video produced by 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 ethos. Nonverbal support is the appearance of the logo of the gym and obese persons telling their story. In general, there is much less focus on experts in the Malaysian videos.

Swedish pathos: Also with respect to *pathos*, in the *Swedish* data the use of overweight persons talking about having fun while training, having found the right diet etc. can evoke sympathy and inspiration. There is also talk about problems, risks, and diseases, which can evoke fear, and talk comparing operation 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 the 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 product introduced. Music is used in all Malaysian videos. In addition, in one video the setting dark grey, the actor's dress dull, the facial expressions showing suffering. On the other hand, showing a happy, healthy family and showing gym activity - trying to evoke inspiration and action.

Conclusions: There are some similarities, but major differences seem to be at work in the two countries, some of the main ones concerning 1) the length and type of videos and 2) the use of logos versus pathos and 3) the choice of multimodal expressions. This is further discussed in the talk.

References :

Allwood, J. (1978). *On the analysis of communicative action*: University of Gothenburg.,
Aristotle, In Kennedy, G. A. (2007). *On rhetoric: a theory of civic discourse*. New York: Oxford Univ. Press.,
Ramage, J. D., Bean, J. C., & Johnson, J. (2015). *Writing arguments: A rhetoric with readings*: Longman.

Identification of Emphasised Regions in Audio-Visual Presentations

Keith Curtis
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
kcurtis@computing.dcu.ie

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

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

I. 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. 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 summarization or browsing of audio-visual content.

Previous work has explored identification of emphasised speech using only 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.

II. Previous Work

Previous research on emphasis detection has focused on using only the speech channel taking the top 1 percentile of pitch values. These were found by expert human annotators to be reliably emphasised. Work by Arons [1] found that detected emphasised areas in the top 1 percentile of pitch values formed a good basis for automatic summarization of speech-only recordings. He et al. [2] 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.

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. [2] 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.

III. Multimodal DataSet

For this investigation we use content selected from the International Speech Conference Multi-modal Corpus (SCMC) from Curtis et al. [3]. This dataset consists of audio-visual recordings of presentations from an academic conference. For this study four presentations were selected form the corpus, two of which had male presenters and two of 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 [3], and to exclude regions of speech not of the presenter.

IV. Multimodal Feature Extraction

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

Pitch: extracted using AutoBi Pitch Extractor [4]. We use default min and max values of 50 and 400 respectively.

Intensity: extracted using AutoBi Intensity Extractor described by Rosenberg [4]. This generated an Intensity contour using default parameters of a minimum intensity of 75dB and a timestep of 100ms.

Head movement: extracted from OpenCV [5] using Robust Facial Detection described by Viola and Jones [6]. 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 head movement by taking the Euclidean distance between pixel points in corresponding frames.

Speaker Motion: extracted using an optical flow implementation in OpenCV [5] described by Lucas and Kanade [7]. We calculated the total pixel motion changes from frame to frame to put more weight on directional changes in motion.

All features were normalised over their entire range for each speaker.

V. Experimental 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. The annotators were asked to mark areas where they consider the presenter to be giving emphasis. There was actually much disagreement between the annotators over areas of emphasis. We consider this due to the

Fig. 1. Presenter: mid-Emphasis

high level of subjectivity on just what it means to emphasise. To better understand the characteristics of regions consistently labeled as emphasised, we studied areas of agreed emphasis between the annotators. It was clear from this analysis, that consistent with earlier work, all agreed 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 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 2 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 top 20 percentile of gesticulation down to the top 40 percentile of values respectively. This resulted in 83 candidate areas of emphasis from our dataset. These candidate regions were each judged for emphasis by three human judges each, with the majority vote on each candidate emphasis region taken as the gold standard label for final agreement of emphasis.

VI. Results and Analysis

Of the 83 candidate areas of emphasis extracted from presentation segments, 18 had pitch values in the top 1 percentile after normalization. Of these 18 candidate areas, 4 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 depends on gesturing and/or other movement in addition to pitch.

15 of the candidate areas of emphasis were in the top 5 percentile of pitch values extracted. 3 of these were accompanied by gesturing on the part of the presenter. All 3 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. 7 were accompanied by gesturing and all of these were judged by human annotators to be emphasised. 26 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 labeled by our human annotators. The meaned 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, 13 additional samples of non-emphasised speech were used. Three additional human annotators were recruited to annotate new candidate emphasis area. 13 of the 15 emphasised areas were selected by annotators as emphasised.

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.

VII. Conclusions and Further Work

Previous work on emphasis detection had looked at the concept in the context of the audio-stream only. Our small initial study shows that emphasis of speech can depend upon speaker gesticulation in addition to pitch. Intensity did not show any significant correlation with emphasis. These results demonstrate the importance of gesturing for emphasis in the audio-visual stream.

VIII. 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.

References

[1] B. Arons, “Pitch-based emphasis detection for segmenting speech recordings.” in *ICSLP*, 1994.

[2] L. He, E. Sanocki, A. Gupta, and J. Grudin, “Auto-summarization of audio-video presentations,” in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 489–498.

[3] K. Curtis, G. J. Jones, and N. Campbell, “Effects of good speaking techniques on audience engagement,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 35–42.

[4] A. Rosenberg, “Autobi-a tool for automatic tobi annotation.” in *INTERSPEECH*, 2010, pp. 146–149.

[5] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. ” O’Reilly Media, Inc.”, 2008.

[6] P. Viola and M. J. Jones, “Robust real-time face detection,” *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] B. D. Lucas, T. Kanade *et al.*, “An iterative image registration technique with an application to stereo vision.” in *IJCAI*, vol. 81, 1981, pp. 674–679.

Preliminary Studies in Laughter and Smile Patterns in Dyadic Conversation for HCI Applications

Kevin El Haddad[1], *Emer Gilmartin*[2], *Hüseyin Çakmak*[1], *Stéphane Dupont*[1], *Thierry Dutoit*[1], *Nick Campbell*[2]

[1] TCTS lab - University of Mons
[2]Speech Communication Lab - Trinity College Dublin
kevin.elhaddad@umons.ac.be, gilmare@tcd.ie

1. Introduction

Conversation, both practical and social, contains a large amount of laughter and smiling behaviour. This behaviour is believed to aid the social bonding vital to human relations, and may contribute to task success in areas such as service encounters, meetings, and education. To more fully understand the dynamics of laughter and smiling behaviour in dyadic conversations, we have performed analyses of the patterning of laughter and smiling in a selection of conversations drawn from the Cardiff Conversation Database (CCDb) [1]. Below we briefly review current knowledge on smiling and laughter, describe the data and methods used in our experiments, and discuss our results and future work.

2. Smiling and Laughter in Spoken Interaction

Laughter and smiling are regarded as basic human social signals, believed to have evolved from primate facial displays - the silent bared-teeth display (smiling), and relaxed open mouth display or play face (laughter). Both signals occur frequently in human interaction, are associated with prosocial affiliation [2], and are thus posited to aid social bonding, and indeed enhance more practical interactions. Smiling, particularly in the case of spontaneous or 'Duchenne' smiles, is believed to signal willingness to co-operate and to engender trust in the recipient. Laughter has been described as predominantly a social rather than a solo activity, is universally present in humans, part of the 'universal human vocabulary', innate, instinctual, and inherited from primate ancestors [3, 4]. It has been described as a social cohesion or bonding mechanism present in primates [5, 2]. It has been suggested that laughter can provide clues to dialogue structure [6]; this has been seen around topic changes where laughter seems to provide an interlude of social bonding and ward off uncomfortable silence [7]. Laughter episodes take a range of forms – from loud bouts to short, often quiet chuckles. It is multimodal, comprising a stereotyped exhalation of air from the mouth in conjunction with rhythmic head and body movement [8, 9]. However, while this stereotypical laugh is generally produced upon asking an informant to laugh, it has been shown to be only one of several manifestations of laughter present in social interaction, and often not the most prevalent [10]. In conversation, laughter generally punctuates rather than interrupts speech, although it can occur within speech as smiled speech or speech laughs. In recent years, there has been increasing research into the occurrence and modelling of laughter in interaction [11, 12, 13], and the audiovisual synthesis of laughter [14]. Human-computer interaction (HCI) incorporating signals such as smiling and laughter could be more acceptable to users, particularly in social domains. Implementation of such signals could be advanced by knowledge of how smiling and laughter is distributed in human-human casual conversation. This is the goal of the work described here.

3. Data

Our experiments are carried out on the Cardiff Conversation Database (CCDb), a multimodal collection of natural dyadic talk comprising 30 interactions recorded audiovisually (2D video and audio). The conversations are manually segmented for speech/silence, and a subsection (8 conversations) are annotated with paralinguistic and non-verbal phenomena, including facial, audio and body gestures expressing states such as agreement, disagreement, and surprise. Annotations were also made for smiles and laughter, making this corpus suitable for our current study as the labels indicated that there was a significant presence of both phenomena. However, we noted that the existing annotations were not suitable for our purposes and thus we re-annotated smiles and laughs from scratch. Our initial explorations were focussed on the 8 fully annotated conversations, and we plan to extend our work to the remaining conversations.

We first annotated the conversations for incidences of smiling and laughter, including speech laughs. We then added information on the intensity of the laughter at 3 levels, and on the intensity of smiles, also at 3 levels. Laughter intensity was obtained subjectively - two annotators scored each of the labelled laughs from 1 to 3, and the average score was used as the intensity level – 1 or lower for low, 2 for medium and 3 for high. The intensity of smiles was based on the width of lip spreading and mouth opening. We then identified and attributed listener and speaker roles throughout the conversations, using the speech labels. We could also determine the smiling/laughter (or neutral) state of each of them using the corresponding labels.

4. Experimental Questions

Two-party conversation is a joint effort, with both participants working together in intricate co-operation to exchange and elaborate information, In the current study, we are interested in how laughter and smiling propagates within and between speakers. We thus simplify the roles of the participants at any time to *speaker* and *listener*. We consider a speaker to be the participant that is speaking, and holds the floor at a particular time. His interlocutor will be considered a listener although he may well be producing short verbal or non-verbal vocalisations in the form of backchannels as the speaker talks.

We are interested in exploring the following questions using the CCDb:

- Do speaker's laughs and smiles influence the production of smiling and laughter by the listener?
- Does the listener cause the smiles/laughs or does the speaker cause them?
- Are the listener laughs always preceded or followed by smiles?
- Is there a relationship between the laughs/smiles intensity levels?

In this paper we present our preliminary work towards answering the above-mentioned questions. We give the characteristics of the newly annotated dataset. Below we describe our work in progress on these topics.

5. Experiments and Results

After reannotation of the smiles and laughs as described previously, the total number of listeners' smiles and laughs instances is 23 and 43 respectively. As preliminary results, we have calculated the percentage of co-occurrences of the listeners' smiles and laughs with the speakers' smiles and laughs. We also calculated the percentage of speakers' smile and laughs preceding the onset of listeners' smiles and laughs by 1 second. The results of these studies are shown in Table 1.

Spk Lis	Sm	L	prec.Sm	prec.L
Sm	18 (41.86%)	11 (25.58%)	13 (30.23%)	8 (18.60%)
L	14 (60.87%)	5 (21.74%)	8 (34.78%)	6 (26.09%)

Table 1: Listeners' smiles (Sm) and laughs (L) frequencies with respect to co-occurring speakers' smiles and laughs. Also with respect to the speakers' smiles and laughs occurring 1 second before the start of the listeners' smiles and laughs (prec.Sm and prec.L respectively).

As we can see it seems that the listeners' smiles and laughs both co-occur the most frequently with speakers' smiles. Also the listener's smiles and laughs are preceded most frequently by the speakers' smiles. We also report in Table 2 the number of smiles and laughs per intensity level, expressed by the listeners and speakers. We can see from this table that the speakers and listeners smiles and laughs distributions per level are similar in this dataset. Concerning the smiles, the distribution seem almost evenly distributed for both speakers and listeners while their laughs seem to be mostly low and middle levels with rare occurrences of high level laughs.

6. Discussion

Several conclusions can be drawn from these preliminary results. First, in this dataset, expression levels seem to vary in the course of the conversation since a significant number of smiles and laughs can be found at different levels. Also, we can conclude that high level laughter is not a common feature in this dataset. We can also note that smiles, along with the content of the speakers' speech, could contribute to making the listeners smile or laugh. Or it could be the opposite. It may be that the listeners' laughs or smiles trigger mostly smiles from the speaker but also laughs. In either case, the presence of a relationship between occurence of social signals in speakers and listeners would lend support to theories of alignment or mirroring of interactive behaviour. We are currently investigating these questions and expect results very shortly. Our goal is to extend this preliminary analysis to the entire CCDb and maybe also to other databases. The results of the work initiated by this paper would allow us to have a better understanding of the occurrences of smiles and laughs in a dyadic conversation. We would then be able to create probabilistic models these expressions occurrences in order to improve HCI applications.

Turn Expression	Lis	Spk
Sm-low	13	19
Sm-med	13	15
Sm-high	17	22
L-low	12	7
L-medium	9	9
L-high	2	2

Table 2: Speakers' and listeners' smiles and laughs frequencies per levels

7. Acknowledgements

This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences and Technologies ERA-NET (CHISTERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot. It was also partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientifique (FNRS), contract no. R.50.01.14.F.

8. References

[1] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, Jun. 2013, pp. 277–282.

[2] M. Mehu and R. I. Dunbar, "Relationship between smiling and laughter in humans (Homo sapiens): Testing the power asymmetry hypothesis," *Folia Primatologica*, vol. 79, no. 5, pp. 269–280, 2008. [Online]. Available: http://www.karger.com/Article/Abstract/126928

[3] R. Burling, *The talking ape: How language evolved*. Oxford University Press, USA, 2007, vol. 5.

[4] R. R. Provine, "Laughing, tickling, and the evolution of speech and self," *Current Directions in Psychological Science*, vol. 13, no. 6, pp. 215–218, 2004.

[5] P. J. Glenn, *Laughter in interaction*. Cambridge University Press Cambridge, 2003. [Online]. Available: http://www.lavoisier.fr/livre/notice.asp?ouvrage=1435545

[6] E. Holt, "Conversation Analysis and Laughter," *The Encyclopedia of Applied Linguistics*, 2013.

[7] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell, "Laughter and Topic Transition in Multiparty Conversation," in *Proc. SigDial*, Metz, France, 2013, pp. 304–308. [Online]. Available: http://acl.eldoc.ub.rug.nl/mirror/W/W13/W13-4045.pdf

[8] J. A. Bachorowski and M. J. Owren, "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context," *Psychological Science*, pp. 219–224, 1995.

[9] D. W. Black, "Laughter," *JAMA: the journal of the American Medical Association*, vol. 252, no. 21, pp. 2995–2998, 1984.

[10] J. Trouvain, "Segmenting phonetic units in laughter," in *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona*, 2003, pp. 2793–2796.

[11] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings." in *INTERSPEECH*, 2007, pp. 1258–1261.

[12] E. Holt, "Laughter at Last: Playfulness and laughter in interaction," *Journal of Pragmatics*, 2016.

[13] J. Ginzburg, Y. Tian, P. Amsili, C. Beyssade, B. Hemforth, Y. Mathieu, C. Saillard, J. Hough, S. Kousidis, and D. Schlangen, "The Disfluency, Exclamation and Laughter in Dialogue (DUEL) Project," in *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt), Posters*, 2014.

[14] H. Çakmak, "Audiovisual Laughter Synthesis - A Statistical Parametric Approach," Ph.D. dissertation, University of Mons, 2016.

TOWARDS CLASSIFICATION OF HEAD MOVEMENTS IN AUDIOVISUAL RECORDINGS OF READ NEWS

Johan Frid[1,2], Gilbert Ambrazaitis[2], Malin Svensson Lundmark[2], David House[3]
[1]Humanities Laboratory, Lund University, Sweden
[2]Linguistics and Phonetics, Centre for Languages and Literature, Sweden
[3]Department of Speech, Music and Hearing, KTH, Sweden
{johan,frid|gilbert.ambrazaitis|malin.svensson_lundmark}@ling.lu.se, davidh@speech.kth.se

This study is part of a project investigating levels of multimodal prosodic prominence, as resulting from an interplay of verbal prosody (pitch accents) and visual prosody (head and eyebrow beats). 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, and 2) to identify labelling-related problems and see if it might be useful for the improvement of the labelling process to have access to movement data.

Our materials consist of Swedish television news broadcasts and comprise speech from four new readers (two female) and about 1000 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.

This corpus was previously 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. Then, ELAN was used to determine for each word if there was head movement or not, where ‘presence’ was defined as the event that the head rapidly changed its position, roughly within the temporal domain of the word. This was done based on the complete audio-visual display, by three annotators independently of each other. Finally, the three annotations were compared, and for the analyses (as well as for the present study), an annotation was counted as such in the event of an agreement between at least two annotators. These annotations constitute the point of departure for the present study. For a more detailed discussion of our definition of beat head movements and our other multi-modal annotations (prosodic prominence), see Ambrazaitis et al (2015).

For the video analysis we used the frontal face detection functions in the OpenCV library 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 equvalent in this context - as coordinates in the x-y plane, as illustrated in Figure 1.

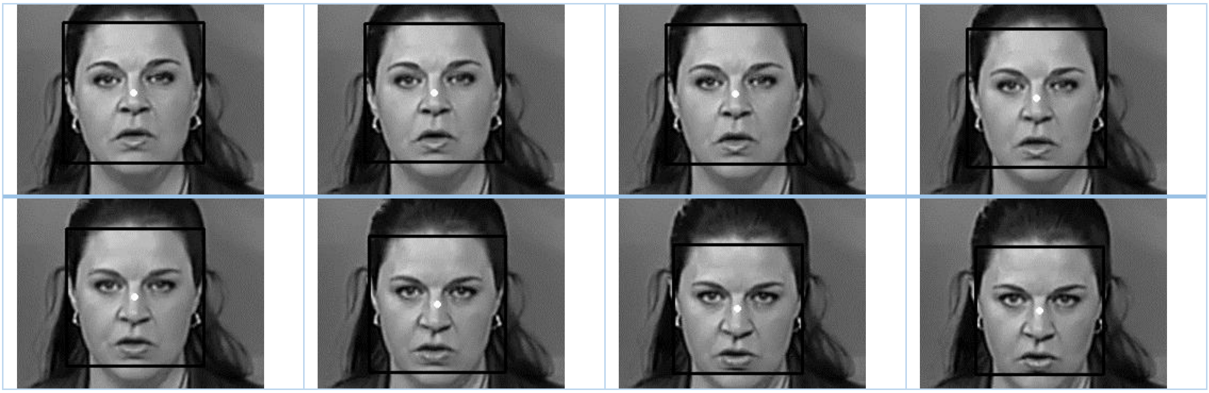


Figure 1. Faces detected in successive frames during a head movement. The black square is the detected face, the white dot (at the center of the square) is the x-y coordinate we use.

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 angular velocity and acceleration are calculated as the Euclidean distance of the x- and y-components. This is shown in Figure 2, where we also show how we can compare the movement functions with the intervals of our word-related head movement labelling.

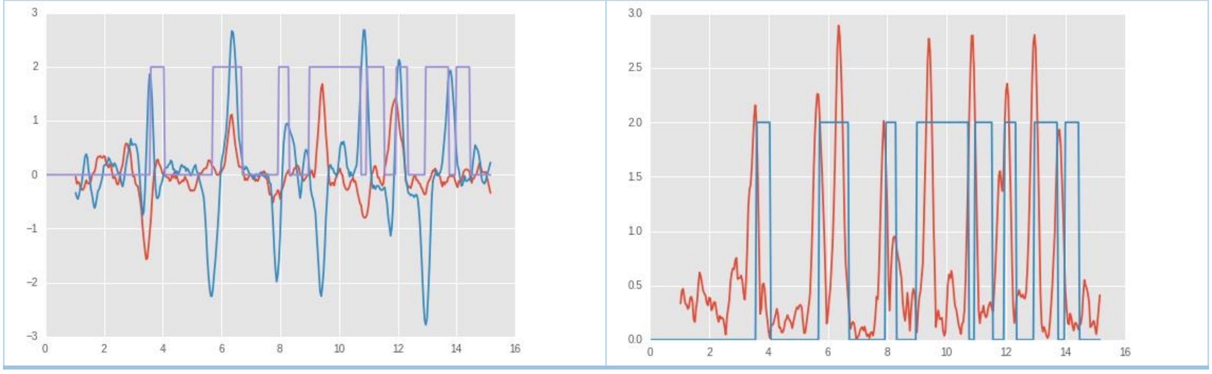


Figure 2. Left: x-velocity (red), y-velocity (blue) and word intervals (purple) as a function of time. Right: angular velocity (red) and word intervals (blue) as a function of time.The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

From each of the six curves (x-velocity, y-velocity, x-acceleration, y-acceleration, angular velocity and angular acceleration) we calculate four features per word: average, max, min and amplitude (max-min). We then trained a classifier by feeding the features into a machine learning algorithm. Our test corpus contains 1047 words. Of these, 818 words are labelled as not having head movement (about 78%), and 229 are labelled as having head movement. We ran a 10-fold cross-validation using xgboost (Chen & Guerstin 2016) and this gave us 85% correctly classified words. We thus perform better than the 'majority' vote, which would have assigned 'no movement' to all words.

The classifier may be helpful for head movement labelling in its own right. Moreover, as may be evident from Figure 2, 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. By visualising the head movements in this way we might indicate that some labels needs to be adjusted. Our goal for future work is to improve the classifier and to integrate this in a tool which could facilitate head movement labelling.

Acknowledgements

This work was supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, 2014–2018; grant number 821-2013-2003) and a grant from the Marcus and Amalia Wallenberg Foundation (grant number 2012.0103).

References

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.

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

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.

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

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

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). 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 of 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 physician's gestures provided unambiguous indications of the relevant body region, but their speech provided necessary information about why it was relevant (e.g., proposing 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 gestures 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, using an inhaler, 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. Gestures fulfilling these criteria 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). Further, demonstrations of actions included postural portrayals of feelings (e.g., sitting up straight suddenly to portray shock).

Our 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 with excerpts from real primary care encounters filmed in the UK (9 physicians and 12 patients). The excerpts illustrated a variety of clinical situations. All 12 excerpts were analyzed, providing approximately 29 minutes of material. Gestures and speech were annotated using ELAN. After locating all gestures, they were categorized as serving one of the following functions: semantic (with concrete or abstract referents; see definitions in Gerwing & Dalby, 2014), beat, or interactive (e.g., Bavelas, Chovil, Lawrie, and Wade, 1992). Concrete semantic gestures were further examined to locate all body-oriented gestures, which were selected for detailed qualitative analysis.

3. Results

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

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. E.g., a gesture might indicate the relevant body part (e.g., the chest), but speech conveyed the sensation (e.g., pain, lack of pain, tenderness, itchiness), the intensity, or time (e.g., duration, past, present). Of the 104 body-oriented gestures, 49/66 (0.74) were direct references to a particular body part (as opposed to a demonstration of an action) in which the patient did not specify the body part in the speech. Sometimes the body part was missing from speech (e.g., to describe where he had a rash, one patient said "it's not just here, it's here, here…" while pointing to his armpits, thighs, and groin). Sometimes, patients mentioned an area 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).

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 21/24 (0.88) were direct references to a particular body part (as opposed to a demonstration of an action) in which the physician did not specify 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. Some further examples of how physicians and patients used body-oriented gestures are the following:

- Physicians and patients displayed their understanding. 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). Note that these gestures did not function to draw the physician's attention to those areas of his chest; rather they displayed the patient's understanding of the instructions, contributing to accomplishing the clinical task of securing mutual understanding regarding his treatment plan. Another physician gestured towards his chest while asking for elaboration about the patient's symptoms, displaying his understanding of the location of the patient's pain. The patient then corrected him by indicating on her own body where he had pointed (while saying "it's not pain there") and then where she actually felt the pain (while saying "it's inside there"). Thus she indicated the location of her symptom, foregrounded against the physician's displayed understanding.
- Patients used body-oriented gestures with speech to contrast past and present. 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 motioned towards her mouth (depicting taking the pills) and then drooped her body over the table (as if in dread) while

smiling. This latter, complex portrayal did more than locate symptoms; instead, the patient used her body and speech to provide an evocative demonstration of the trajectory 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 humor with her facial display.

- Physicians used their own body as well as the patient’s body to keep their speech visibly oriented to the relevant body regions. For example, while explaining systemic effects of eczema, one physician drew circles over the patient’s hand, where the patient had a rash, but then, connecting this symptom to irritation in the patient’s ear while providing some visible cohesion to her earlier explanation of eczema, the physician motioned towards her own ear, where she had previously gestured while explaining the reason for the irritation.
- Physicians also used gestures towards the patient’s body to demonstrate their clinical activities. One physician, 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/implications

Physicians and patients integrated body-oriented gestures with their speech in sophisticated and systematic ways. By using gestures to show locations of symptoms, physicians and patients could refer unambiguously to the relevant body region without necessarily using terminology that was potentially ambiguous. Although these gestures were oriented around the body, they did far more than simply convey information about symptom location; both patients and physicians used them to accomplish complex clinical communication tasks. Further research using the analytical framework is being conducted on interactions between specialists and patients in a Norwegian hospital and on triadic, interpreted primary care encounters filmed in the UK.

5. References

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15,* 469-489.

Bavelas, J. B., & Chovil, N. (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, G. (2011) Elements of meaning in gesture. The Netherlands: John Benjamins B.V.

Gerwing, J. & Dalby, A. M. (2014) Gestures convey content: An exploration of the semantic functions of physicians' gestures. *Patient Education and Counseling, 96,* 308-314.

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

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

Kendon A. (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.

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

Rowbotham, S., Holler, J., Lloyd, D., & Wearden, A. (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.

Adaptive Behavior Generation for Multimodal Child-Robot-Interaction

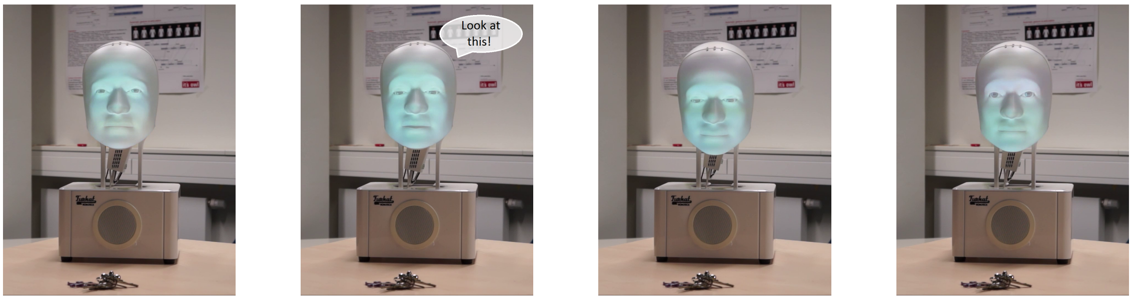
Jacqueline Hemminghaus & Stefan Kopp
jhemming@techfak.uni-bielefeld.de, skopp@techfak.uni-bielefeld.de
Social Cognitive Systems, CITEC, Bielefeld University, Germany

Human-Robot-Interaction is an increasingly studied field with many disciplines involved. Dependent on the different target groups different challenges have to be solved. While robot platforms for elderly people have to deal with their lacking visual and hearing abilities, robots for children have to deal with their developing behavior and communication skills. Additionally, children are easily distracted by anything in the room which might be interesting for them at the moment [1]. Therefore robots should be able to interpret the child's behavior including cues of, e.g., attentive or affective states. Additionally the robot should be able to grab and guide a child's attention to important points in the task space. Depending on the robot platform, different behaviors are possible for this, with different affordances and limitations. As an example, vehicle robots can move around and thus can grab the attention through wobbling or moving back and forth in front of the desired object [2]. Humanoid robots are able to communicate through their verbal or non-verbal behavior [3].

In this abstract we present the development of a social tutoring robot, which can interact with children by interpreting and naturally reacting to the child's behavior. Its main task is to support the development of social and communicative skills in an educative scenario, where mainly two children will play a game together with the robot. The robot acts as a moderator or guides through the game tasks which the children should more of less solve on their own. To minimize the complexity of the setup, the game is displayed on a touch table display which is positioned between the children and the robot. Additionally, sensors, e.g. a Kinect camera, are placed in front of the children to observe where or what each child is attending to and to enable detection of their emotional or motivational state.

As interaction partner we use the Furhat robot head [4], which features gaze, facial expression, head movements and speech. Other robot platforms are possible as well. To model the main interaction process during the game task we use the event based system IrisTK [5] which comes with Furhat and provides modules for multimodal input and interaction management. Because there are only low level behaviors built in, like manipulating the gaze position, facial expression and speech synthesis, we are presently developing a module that can plan multimodal high-level behavior, like attention grabbing or motivating. These behaviors should be adaptive to the communicative affordances of any the robot used. To simplify the realization of these behaviors we extend IrisTK with parts of the Articulated Social Agents Platform [6], which provides modules to schedule and realize synchronized behavior for different output platforms.

As a starting point for the behavior planning we focus on the attention management process within a simplified setting with only one user. This task can be split into two parts. First, Furhat has to react to the attention guidance of a user. Therefore, Furhat has to check whether it looks at the same object than the user by switching its gaze between both, user and target object [7]. Secondly, it has



(a) Facial Gesture (b) Speech (c) Gaze at object (d) Gaze at person

Figure 1: Possible behaviors of the robot to guide the attention of the user to the keys in front of Furhat.

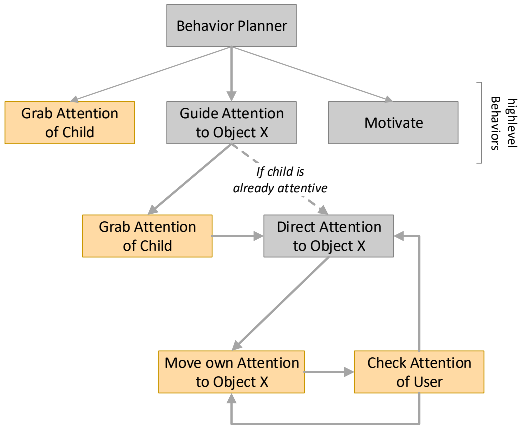

Behavior Planner
Grab Attention of Child
Guide Attention to Object X
Motivate
highlevel Behaviors
If child is already attentive
Grab Attention of Child
Direct Attention to Object X
Move own Attention to Object X
Check Attention of User

Figure 2: Hierarchically order of behaviors executable by the behavior planner. Arrows represent the structure of the execution process. Highlighted nodes indicate a leaf node, which execute one or a combination of base behaviors. As an example the guide attention behavior is split in its subbehaviors to visualize the different phases. As can be seen the grab attention behavior is visualized two times. On the one hand it can be a main behavior to get the attention to the robot itself, on the other hand it's a part of another high level behavior. This shows the reusability of a behavior node.

to guide the attention of the user. Here Furhat may use several modalities to gain the attention of the user and guide it to an object (see figure 1 for an example). First of all, it can gaze at the target object. Additionally it can point out its concern by using facial expression, e.g. raising its brows. If the human is still not responding nor looking at the target object, Furhat can use head turns or more expressive head gestures, e.g. nodding in the direction of the target. Finally Furhat can point out its desire by using speech. Our eventual goal is to enable the robot itself to explore for each task which combination of behaviors, and for each behavior, which combination of modalities, achieves the best attention guidance results under different noisy conditions and for different children.

Our first step is the development of a behavior planner which creates behaviors for different tasks, e.g. attention management or motivating. As can be seen in figure 2, the behavior planner presumes different high-level behaviors, each with a hierarchical structure representing the different tasks of the behavior. If necessary a subbehavior can be further refined as well, until a leaf node is reached, which selects and executes the respective actions. Each leaf node decides by itself which gesture combination should be executed by Furhat. We start with predefined base gestures, which can be parameterized and combined to more complex behaviors, e.g. speech refers to Furhat or an object. A Q-learning algorithm weights each behavior, depending on the responsiveness of the child and the effectiveness of the executed gesture(s).

References

[1] T. Salter, I. Werry, and F. Michaud, "Going into the wild in child-robot interaction studies: Issues in social robotic development", *Intelligent Service Robotics*, 2008.

[2] J. Fink, S. Lemaignan, P. Dillenbourg, P. Rétornaz, F. Vaussard, A. Berthoud, F. Mondada, F. Wille, and K. Franinović, "Which robot behavior can motivate children to tidy up their toys?: Design and evaluation of ranger", in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014.

[3] E. J. Van Der Drift, R.-J. Beun, R. Looije, O. A. Blanson Henkemans, and M. A. Neerincx, "A remote social robot to motivate and support diabetic children in keeping a diary", in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 2014.

[4] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction", in *Cognitive behavioural systems*, 2012.

[5] G. Skantze and S. Al Moubayed, "Iristk: A statechart-based toolkit for multi-party face-to-face interaction", in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012.

[6] S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier, "An architecture for fluid real-time conversational agents: Integrating incremental output generation and input processing", *Journal on Multimodal User Interfaces*, 2014.

[7] N. Pfeiffer-Lessmann, T. Pfeiffer, and I. Wachsmuth, "An operational model of joint attention-timing of gaze patterns in interactions between humans and a virtual human", in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[8] G. Gordona, S. Spauldinga, J. K. Westlunda, J. J. Leea, L. Plummera, M. Martineza, M. Dasa, and C. Breazeala, "Affective personalization of a social robot tutor for children's second language skills", 2015.

Gestures become more informative when communication is unsuccessful

Marieke Hoetjes
Centre for Language Studies
Radboud University, Nijmegen, the Netherlands

It is known that when people produce repeated references to an object during a conversation, that these repeated references are often reduced, for example with regard to the number of words (e.g. Clark & Wilkes-Gibbs, 1986), their acoustics (Bard et al., 2000), but also with regard to the number of gestures (e.g. Galati & Brennan, 2014; Masson-Carro, Goudbeek, & Krahmer, 2014). The reduction process that has been found in repeated references is presumably due to the fact that when repeated references are produced there is usually a context of common ground (Clark & Brennan, 1991). When there is common ground between speakers, a lot of information can be considered given, and this information does not need to be repeated for communication to remain successful. In line with this, previous work (Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015) has shown that although gestures in repeated references may be reduced in some respects, when these reduced gestures are presented to addressees, as much information can be inferred from them as from gestures produced in initial references.

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. Some previous work has been done on this matter (Hoetjes, Krahmer, & Swerts, 2015; Holler & Wilkin, 2011), and has shown that when a speaker receives negative feedback, gestures in the following references are not necessarily reduced, but can increase in size and precision. In the current study, we investigate whether this also means that 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 is done with the addressee in mind.

Sixty-nine participants (21 males, *M* = 21 years old, range 18-28 years old) took part in this study. Participants were presented with 88 short video clips taken from previous work on gestures in unsuccessful communication in which participants had to repeatedly describe complex objects (Hoetjes, Krahmer, et al., 2015). These video clips were presented without the original speech, and each video clip showed a gesture about the shape of one of the objects, that was either produced before any feedback was given, during an initial reference to the object, or after negative feedback was given, in a second or third reference to the object. The task was to match the gesture in the video clip to the correct object: the one during whose description the gesture was produced. Participants were shown 2 objects to choose from: the correct object and a similar looking, but incorrect, object. We counted the number of times that participants chose the correct object, across initial, second and third references and analysed whether these percentages differed from chance level (50%), see table 1[1]. We found that the distribution differed from chance level, $\chi^2(5) = 83.49$, $p < .01$. The percentage of correct answers increased for each repeated reference, that is, after each instance of negative feedback, $\chi^2(2) = 63.02$, $p < .01$.

Table 1. Percentage of correct answers, across conditions (2nd and 3rd references were produced after negative feedback).

	Visibility	No visibility	Total
1st reference	51,6	54,8	53,2
2nd reference	53,7	56,7	55,2
3rd reference	59,2	61,8	60,5

The results show that participants are better at deciding which object a gesture is based on when this gesture is produced after negative feedback. Previous work has already shown that gestures in repeated references are not always reduced, and can increase in size and precision after negative feedback. However, it was unclear whether these changes in gesture production are also noticeable and useful for the addressee. We can now provide evidence that gestures that are produced when communication is unsuccessful also become more informative for the addressee.

This study suggests that gestures can provide valuable information in a discourse context. In this case, participants were able to pick the correct object above chance level, after only viewing one gesture (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 keep the addressee in mind, and thereby help to keep the overall communicative situation as successful as possible.

Acknowledgments

Many thanks to Malu Hanssen for 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.

[1] Table 1 also shows results for the independent variable visibility (i.e. whether the gesture was produced in a context of mutual visibility between the speaker and addressee, or not), which we show for completeness' sake, but is not the focus of the current abstract.

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.

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.

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2014). On the automaticity of reduction in dialogue: Cognitive load and repeated multimodal references. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Quebec City: Cognitive Science Society.

Gaze distribution during fillers: empirical data on the difference between Dutch ‘euh’ and ‘euhm’

In natural conversations, people often use fillers such as ‘euh’ and ‘euhm’. These fillers do not always stem from a cognitive production difficulty but can also serve a communicative goal (Lake et al., 2011; Walker et al., 2014). Different interactional functions have been attributed to fillers, such as the marking of discourse boundaries (Swerts 1998), warning for a delay (Clark, 1994), a turn holding, taking or yielding function (Beattie, 1983; Clark & Fox Tree, 2002; Cook & Lalljee, 1972; Maclay & Osgood, 1959). This study explores the correlation between fillers and gaze aversion by a speaker. Just as fillers, gaze aversion can serve different communicative goals, which seem to be related to those of the fillers. For example, people look away when they want to continue their turn (Hirvenkari et al., 2013) and when they start a new turn (Oertel et al., 2012).

More specifically, speakers’ gaze patterns when producing two different manifestations of the filler (‘euh’ vs. ‘euhm’ in Dutch) are compared. The distribution of gaze just before, during and after ‘euh’ and ‘euhm’ was studied in three triadic conversations between acquainted students. In each triad, the participants had a free conversation of approximately 15 minutes. The triads were recorded in an experimental setting, using mobile eye tracking glasses (Gullberg & Kita 2009; Jokinen 2010; Brône & Oben 2015). After synchronizing the videos made by the mobile eye trackers and the static camera, the conversations were transcribed and annotated for gaze target and fillers.

We expect a difference in gaze pattern between ‘euh’ and ‘euhm’, as the two realizations of the filler also differ in other aspects. As already observed before, the formal realization of a filler affects its communicative function. The vocal ‘uh’/’euh’ seems to warn the interlocutor for a short interruption, whereas the vocal-nasalic ‘um’/’euhm’ announces a longer delay (Clark, 1994). This functional difference correlates with differences on the formal level. For example, ‘euhm’ occurs more frequently at major discourse boundaries and is more likely to be surrounded by silent pauses than ‘euh’ (De Leeuw, 2007).

The analysis of four triadic conversations reveals that also in gaze distribution, ‘euh’ and ‘euhm’ show some interesting differences. In the data, 81 occurrences of ‘euh(m)’ were found. More than ‘euh’, ‘euhm’ is accompanied by a gaze aversion away from the previous target or gaze fixation to the background. During ‘euh’, on the contrary, speakers’ gaze stays more frequently fixed on an interlocutor. In addition, other interactional differences could be found: ‘euhm’ is surrounded by more silent pauses and occurs more often in turn-final and turn-initial position. The multimodal analysis we present in this study thus supports previous findings of an interactional difference in vocal and vocal-nasalic fillers.

References

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

Brône, G., & Oben, B. (2015). InSight Interaction. A multimodal and multifocal dialogue corpus. *Language Resources and Evaluation,* 49(1), 194-214.

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

Clark, H. H., & Fox Tree, J. E. (2002). Using 'uh' and 'um' in spontaneous speaking. *Cognition, 84,* 73-111.

Cook, M., & Lalljee, M. G. (1972). Verbal substitutes for visual signals in interaction. *Semiotica*, *6*, 212-221.

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

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behaviour,* 33, 251-277.

Hirvenkari, L., Ruusuvuori, J., Saarinen, V.-M., Kivioja, M., Peräkylä, A., & Hari, R. (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PLoS ONE*, *8*(8), e71569.

Jokinen, K. (2010). Non-verbal signals for turn-taking & feedback. *Proceedings of the 7th International Conference on Language Resources & Evaluation* (LREC)*, 2961-2967.*

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

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, *15*, 19-44.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze Patterns in Turn-Taking. In *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012, Vol. 3* (pp. 2243-2246). Portland, Oregon.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30, 485-496.

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

Kinesic code-switching among bi-culturals: A multimodal analysis

Purpose

The purpose of this study is to find out if people who belong to two cultures (bi-culturals) and have two first languages also change their kinesics when they switch from speaking in one language to another. Is there a deeper connection between the communication systems behind such a change? Goldin-Meadow (2003) and McNeill (2000) suggest that language and gestures go together as one communication system. A language used in an expressive culture should be accompanied by vivid gestures while a language used in a reserved culture should be accompanied by moderate gestures. This study can support or question the one-system hypothesis.

Background

The one-system hypothesis (Goldin-Meadow, 2003; McNeill & Duncan, 2000) suggests that words and gestures go together and function as a unison system. A faster speech rate would probably correlate with a higher velocity in hand movements. A two-system hypothesis (Krauss, 1999) suggests that gestures support the verbal system and that gestures can become more intense when the speaker is trying to find the proper words. A lower speech proficiency could increase gesturing.

To test these hypotheses we selected participants that are fluent in two languages and belong to two cultures. We wanted a mix of cultures that are obviously different. One way to differentiate cultures is to use one of the several cultural dimensions that Hall (1969; 1977) or Hofstede (Hofstede, Hofstede & Minkov, 2010) have suggested and pick cultures from different ends. The most relevant dimension for the present study is Matsumoto and Hwang's (2013) distinction between expressive cultures and reserved cultures. People who belong to expressive cultures gesture more, touch more, stand closer, speak louder and faster, and seek more eye contact. This description is similar to Hall's contact cultures versus non-contact cultures (see also Ting-Toomey & Chung, 2005).

Method

The participants in the study identify themselves both with Swedish culture (a reserved culture) and Mozambican culture (an expressive culture). They speak Swedish and Portuguese fluently. The criteria for being selected as a participant is to have lived in both countries and to have spoken both languages for at least five years (all of the participants surpassed this by far). Seven individuals, three men and four women, participated in four test conditions each with the same interviewer:

- An interview in Portuguese, face-to-face (10 minutes)
- An interview in Swedish, face-to-face (10 minutes)
- An interview in Portuguese, audio only (9 minutes)
- An interview in Swedish, audio only (9 minutes)

The reason we included the audio only condition was to avoid any mirroring effects between interviewer and interviewee. If there are differences in both face-to-face and in audio only between languages, the outcome is more valid.

The full test was recorded at 179 frames per second by an eight-camera motion capture system. Every participant, including the interviewer, had 21 markers on strategic body parts to measure gesture activity.

Result

After the statistical analysis (the motion capture system generates over 100,000 data units per marker and test condition), it is very clear that the participants move their hands and heads more when they speak Portuguese compared to when they speak Swedish. There is also an interesting asymmetry in their hand movements when the participants speak Portuguese. They move their right hand significantly more than their left. When the participants were speaking Portuguese the head marker and both hand markers were highly correlated, meaning that they moved head and hands in synchrony. There seems to be a cultural movement pattern that differentiates the Mozambican identity in the participants from the Swedish identity in the participants. The difference is smaller but significant in the audio only condition. This supports the hypothesis that spoken language and gestures go together. There are no reasons to believe that the increase in movement while speaking Portuguese is caused by lower language proficiency since six of the participants were born in Mozambique and learned Portuguese first or together with the Swedish language. The participants did not gesture more to find Portuguese words.

Literature

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge: Harvard University Press.

Hall, E. T. (1969). *The hidden dimension*. New York: Anchor books.

Hall, E. T. (1977). *Beyond Culture*. New York: Anchor books.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations. Software of the mind*. New York: McGraw-Hill.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science* 7, 54–59.

McNeill, D. (ed.) (2000). *Language and gesture*. Cambridge: Cambridge University Press.

McNeill, D., & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In McNeill, D. (ed.) *Language and gesture*. Cambridge: Cambridge University Press.

Matsumoto, D., & Hwang, H. S. (2013). Cultural influences on nonverbal behavior. In D. Matsumoto, M. G. Frank & H. S. Hwang (eds.), *Nonverbal communication. Science and applications*. London: Sage.

Ting-Toomey, S., & Chung, L. C. (2005). *Understanding intercultural communication*. Los Angeles: Roxbury Publishing Company.

Title: Kinesic code-switching among bi-culturals: A multimodal analysis

By: Mikael Jensen and Linnea Moreira Emanuelsson

University of Gothenburg

Department of Applied IT

(Unit for Cognition and Communication)

Forskningsgången 6

412 69 Göteborg

mikael.jensen@gu.se

031-722 2342

Laughing and co-construction of common ground in human conversations

Katri Hiovain ja Kristiina Jokinen
University of Helsinki and University of Tartu
kristiina.jokinen@helsinki.fi

ABSTRACT

This paper focuses multimodal human-human interactions, and describes our studies on how participants are engaged in conversations, and create rapport through laughter and shared knowledge. We examine the video data from Finnish, Estonian and North-Sami conversations, and compare how the participants' multimodal behavior indicates their experience and engagement in interaction. In particular, we study laughters, body movements, and speech contributions, and how they indicate the participants' engagement in conversations. The Finnish and Estonian data share the same conversational activity setting (i.e. first-encounter dialogues), and thus offer a starting point for interesting intercultural studies. As for the North-Sami data, the activity type is different so direct correlation is not possible, but it will be possible to juxtapose North Sami conversational speech and types of laughter with the two other languages so as to set out modelling of conversational speech for North Sami.

1. INTRODUCTION

Laughter is usually related to joking and humour, but it has also been found to occur in various socially critical situations where its function is connected to creating social bonds as well as signalling relief of embarrassment. While laughing is an effective feedback signal that the participants use to show their benevolent attitude, it can also be used as a subtle interaction strategy to distance oneself from the partner and from the discussed topics, i.e. as an acceptable way to disassociate oneself from the conversation. Lack of laughter is, consequently, associated with serious and formal situations where the participants wish to keep distance in their social interaction. Following our earlier preliminary studies on laughter in North-Sami conversations (Hiovain and Jokinen 2016), we assume that laughing is a social signal that regulates the creation of common ground among the participants and reinforces rapport and the feeling of togetherness in situations which are new and confusing.

This paper focuses on laughing and its function in conversational video corpora collected in three linguistically related communities: Finnish, Estonian and North-Sami. It is tempting to assume that if the linguistic and geographical differences are small, the languages behave in more similar ways than culturally more distant ones, confirming the hypothesis that linguistically and culturally more similar languages share the means to control and coordinate the discussion. While the goal of the study is not to thoroughly answer the question of cultural and linguistic correlations between the particular Finno-Ugric languages, it aims to set the languages side by side so as to study the phenomenon, laughter as a multimodal signal, and its realisation in the different languages in order to shed light on the various functions of laughter in general.

The paper is structured as follows. We first give a brief overview of the previous research on laughter in conversation in Section 2, and then describe the data used in the analysis in Section 3. The analysis is presented in Section 4 and conclusions and future work discussed in Section 5.

2. LAUGHTER IN CONVERSATION

Our earlier work on laughter focussed on North-Sami interactions and describing acoustic properties of different laughter types. We divided the laugher bouts in two types following the earlier studies (Nwokah et al. 1999; Kohler 2008; Tanaka and Campbell 2011): free laugh 'fl' and speech-laugh (laughing speech) 'st', and then further classified the occurrences into 'm' – mirth, 'e' – embarrassed, 'b' – breath, 'p' – polite, 'd' – derision, and 'o' – other. The type 'r' – relief was also searched for in the corpus, but none of the laughters seemed to represent that type. 97% of all laughter types represent the types 'm', 'b' or 'e'.

Concerning the structure of interaction, Vöge (2010) discusses two different positioning of laughter: the same-turn laughter, i.e. the speaker starts to laugh first, or the next-turn laughter, i.e. the partner laughs first. The same-turn laughter shows to the other participants how the speaker wishes their contribution to be taken and thus allows shared ground to be created. Laughter in the second position, however, is potentially risky as it shows that the partner has found something in the previous turn that caught their mind and is laughable; this may increase the participants' disaffiliation, since the speaker might not have intended that their contribution had such a laughable connotation, and so the speakers have to restore their shared understanding. We also checked how the different laughter types occur in the interaction structure: laughter can be regarded as a paralinguistic signal which functions as a means to elicit feedback (by the speaker) and to give feedback (by the partner). These functions are used by the participants to construct their togetherness in the conversations: through eliciting and giving feedback via laughter, the participants reinforce their mutual bond, share embarrassment and uncertainty, and show good humour towards each other.

3. DATA

We will set to investigate how laugh types are used to control and coordinate turn-taking and feedback behaviour, and allow smooth flow of information in conversation. The data is from the Finnish and Estonian First Encounter dialogue corpus (see e.g. Paggio et al. 2010; Allwood et al. 2012; Jokinen & Tenjes, 2012) and from the North Sami Conversational Corpus (Jokinen & Wilcock 2014). The data is transcribed and annotated using Anvil and Praat, and for each laughter bout, its pitch, duration and intensity were extracted of laughter bouts representing every laughter type.

The amount of laughter occurrences in different conversations differs strongly. The analysis of different types of laughter shows some connections to how well participants know each other, how nervous they are, and what kind of relationship they have with each other. We conclude that laughing has several functions that range from a relief burst of embarrassment, to a face-saving means of being polite, and to a specific strategy to distance oneself from the utterance content and to show to the partner that one wants to build rapport and good will. The analysis will reveal the distribution of different laughter types in the coordinating feedback functions and we also study if these functions differ from each other in their acoustic properties. We aim to answer the questions such as:

- Is there a difference in the encoding of laughter between eliciting and giving feedback with respect to F0 and other related parameters?
- Is there a difference in the laughter types between eliciting and giving feedback
- Is there a difference in the encoding of laughter between speakers of different linguistic/cultural backgrounds with respect to F0 and related parameters?
- What is the quantity of laughter, relative to the quantity of speech in different languages?
- How do the participants affect each other in their use of laughter and speech in the different languages?

4. RESULTS

The full paper will show the results of comparing the laughter bouts and their distribution and correlation between the different languages.

5. CONCLUSIONS

This paper studies laugher occurrences in conversational data in three different languages, Finnish, Estonian, and North Sami, and explores how the laughter differs in them and what kind of impact it has on the speakers' experience and engagement in the conversation. The work contributes to our understanding of the interlocutors' engagement in conversational situations and how their experience of the smooth communication is related to laughing and other multimodal signals in interaction. The study is based on laughter analysis in three linguistically related languages, bringing forward interesting possibilities for intercultural comparison of the function and distinctive acoustic features of the laughter bouts in the different but closely related language communities. From the point of view of interactive system design, the models can be used in adapting the system conversational capability to appropriate laughter behaviour, and experimenting with the setups for creating mutual bonds between the interlocutors by associating and disassociating oneself with the interlocutors.

6. ACKNOWLEDGMENTS

We would like to thank the annotators and the participants in the video files as well as the projects supported by Academy of Finland and Estonian Science Foundation.

7. REFERENCES

[1] Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta and P. Paggio (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. In Martin, J.C., Paggio, P., Kuenlein, P., Stiefelhagen, R. and Pianesi, F. (Eds.) Multimodal Corpora for Modelling Human Multimodal Behaviour. Special issue of the International Journal of Language Resources and Evaluation, 41(3-4), 273-287. Springer Online: http://www.springerlink.com/content/x745801041m52553/fulltext.pdf

[2] Hiovain, K., Jokinen, K. 2016. Acoustic Features of Different Types of Laughter in North Sami Conversational Speech. Proceedings of the LREC-2016 Workshop *Just Talking.*

[3] Jokinen, K. 2014. Open-domain interaction and online content in the Sami language. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2014)*, 2014.

[4] Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I., Wilcock, G. 2016. DigiSami and digital natives: Interaction technology for the North Sami language. In: Proceedings of Seventh International Workshop on Spoken Dialogue Systems (IWSDS 2016). Saariselkä.

[5] Jokinen, K., Tenjes, S. 2012. Investigating Engagement Intercultural and technological aspects of the collection, analysis, and use of Estonian Multiparty Conversational Video Data. Proceedings of the Language Resources and Evaluation Conference (LREC-2012). Istanbul, Turkey.

[6] Jokinen, K., Wilcock, G. 2014. Community-based resource building and data collection. In: Proceedings of 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU-2014), pp. 201–206. St. Petersburg

[7] Paggio, P., Allwood, J., Ahlsén, E., Jokinen K., Navarretta, C. 2010. The NOMCO Multimodal Nordic Resource – Goals and Characteristics. In *Proceedings of LREC'10,* Valetta, Malta: ELRA.

Automatic recognition of head movements using velocity, acceleration and jerk

Bart Jongejan, Costanza Navarretta, Patrizia Paggio
CST-University of Copenhagen
Njalsgade 140-142
DK-2300 Copenhagen S
bartj@hum.ku.dk

Research strategies for automatically identifying head movements, especially nods and shakes, from videos, have either used optical motion capture (Zhao et al., 2012), or have tracked interest points of the face, and then applied machine learning to detect the appropriate movements (Morency et al., 2005; Tan and Rong, 2003). Others have added prosodic and lexical features (Nguyen et al., 2012).

Previous work has shown that using physiological characteristics of head movements in terms of velocity and muscular forces (acceleration) provided an interesting and promising framework for the automatic annotation of head movements (Jongejan, 2012). The automated annotations correlated well, albeit with a systematic anticipation of a few frames, with manual annotations of the same movement at movement onset; manual annotations, however, were often considerably longer. The algorithm tended to find many small movements rather than long complex ones. Subsequent experiments with the same system (Jokinen and Wilcock, 2014) showed that it was good at detecting some movement types (*nods* and *turns*), and less good at finding others (*up-nods*). Also in this study, however, it is noted that the number of false positives is too large.

The approach to head movement identification described above was based on the first and second derivatives with respect to time of the position of the face: velocity and acceleration. In this work we add the third derivative, called *jerk* in physics. While velocity is change of position per unit of time, and acceleration is change of velocity per unit of time, jerk corresponds to the change of acceleration per unit of time. The expectation is that a sequence of frames with a high value for jerk in the horizontal or vertical direction is indicative of the most effortful and dynamic part of the movement (the part that researchers have called *stroke* (Kendon, 2004), or *apex* (Loehr, 2007)).

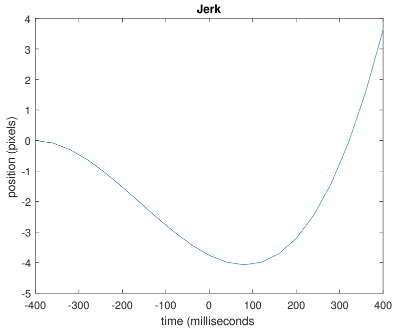


Figure 1: Jerk in an idealized nod. The figure depicts the relative position of the head in a time period of 800 milliseconds. Initially, at -400 ms, the head is at rest in position 0. Due to a negative (downward) acceleration it moves a few pixels down, but the positive, constant (43.74 $pixels/s^3$) 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.

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 idealized 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 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 upward.

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. 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 upward trajectory, the movement proceeds downward and then up just as in Fig. 1.

Based on the values of velocity, acceleration and jerk, the relevant video frames have been classified using the vector support classifier implemented in the LibSVM software (Chang and Lin, 2011) as belonging to various movement types. The use of machine learning is also a new feature of the present approach, compared to the one in (Jongejan, 2012), where thresholds had to be manually set by the user in order for the system to distinguish between presence and absence of

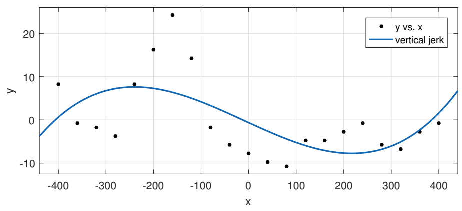


Figure 2: Jerk during a nod computed from data produced by the ANVIL Facetracker. The figure depicts the relative vertical position of the head in a time period of about 800 milliseconds, or 21 frames (assuming a frame rate of 25 frames/s). Points on the curve approximate the relative position of the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

movement.

The head movement recognition software is implemented as a plugin to the ANVIL annotation software tool (Kipp, 2004) and using OpenCV (Bradski and Koehler, 2008).

Velocity, acceleration and jerk are computed for every frame, both horizontally and vertically, giving a total of six independent values. Since a single frame is not sufficient for the computation of any of these physical quantities (only the current position is reported by OpenCV), for velocity we include the previous three and next three frames in the computation, giving a total of seven frames. For acceleration and jerk we use even more frames to reduce the effect of noise in the data to an acceptable level, 14 and 21 frames, respectively.

Including many more frames than the minimally required number[1] effectively smooths out noise, but also head movements of short duration that quickly follow one after another. For example, some of the 'outliers' in Fig. 2 might indicate that the person made several head movements during the 800 ms window. At the cost of a noisier signal, a lower number of frames could be chosen in order to increase the resolution of movements of short duration. The frame-wise data for velocity, acceleration and jerk are stored in three ANVIL annotation tracks that can be seen near the bottom of the annotation window, see Fig. 3. Each annotation cell in each of those rows contains point data that allow ANVIL to visualise the current velocity, acceleration or jerk, as illustrated in Fig. 4, which shows the ANVIL video window while the participant to the left is making a nod. The arrow indicates jerk direction and strength.

The ANVIL annotation was used as input to a number of SVM classifiers. For all the classifiers, each data point (a line in the input file) corresponds to a frame, and contains velocity, acceleration and jerk values. In addition, it contains a binary feature indicating whether the frame belongs to head movement or not, and the head movement class, which is one of 'head backward', 'head forward', 'nod', 'shake', 'side turn', 'tilt', 'up-nod', 'waggle', 'head other' and 'none'. We also added a feature indicating whether the movement is horizontal, vertical, or 'other'. After training and testing the SVM classifier, the predicted head movements were copied to the ANVIL file in a new annotation track which can be seen above the bottom three tracks in Fig. 3.

Head movement	movements	frames	accuracy %	min/ max %
(none)	128	6277	94.87	33/100
Waggle	2	30	50.00	31/65
HeadOther	27	423	36.17	0/82
Tilt	24	560	27.68	0/88
Up-nod	8	69	27.54	0/60
Nod	12	235	24.68	0/85
SideTurn	44	1111	21.60	0/89
Shake	14	292	15.41	0/44
Head-Backward	11	248	13.31	0/53
Head-Forward	12	170	11.76	0/38

Table 1: Frame-wise detection of different head movement types. Data obtained using a binary SVM classifier which predicts the presence of a head movement based on velocity, acceleration and jerks measurements. The prediction presence/absence of movement is shown with respect to the various head movement types, which this classifier does not distinguish among. "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 for all movements of a given type. "Min/max" are the minimum and maximum ratios found for all movements of a given type.

Classifier	Avg accuracy Margin 0	Avg accuracy Margin 17
Binary	71.09	73.47
HVO	70.42	74.13
All classes	70.19	73.64

Table 2: Average accuracy results (over all movement types) for three different classifiers, which distinguish between presence and absence of movement (Binary); horizontal, vertical and other movement (HVO); all different head movement classes (All classes). Margin 0 and Margin 17 refer to the number of non-movement frames included in the predicted movement.

The highest accuracies, shown in Table 1, were obtained with the binary classifier. To reach these accuracies, varying numbers of non-movement frames (margins) were included in the predicted movements. Experiments have shown, in fact, that the optimal margin for bridging automatic annotations is about 17 frames, heightening the overall accuracy by two percent (Jongejan et al., 2016). Almost as good (and

[1] To compute velocity, acceleration and jerk one needs position data from minimally two, three and four frames, respectively.

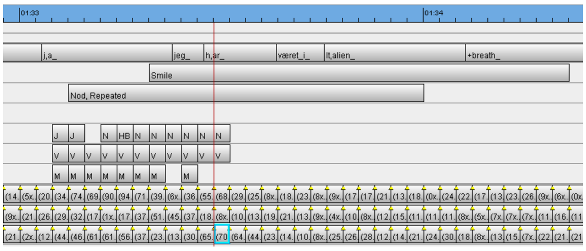

01:33
01:34
j,a_
jeg_
h,ar_
været_i_
lt,alien_
+breath_
Smile
Nod, Repeated
J J N HB N N N N N N
V V V V V V V V V V V
M M M M M M M M
(14. (5x. (20. (34. (74. (69. (90. (94. (71. (39. (6x. (36. (55. (68. (29. (25. (8x. (18. (23. (8x. (9x. (17. (17. (21. (13. (18. (0x. (24. (22. (17. (13. (23. (26. (9x. (6x. (0x.
(9x. (21. (26. (29. (32. (17. (1x. (17. (37. (51. (45. (37. (18. (8x. (10. (13. (19. (21. (13. (9x. (4x. (10. (8x. (12. (15. (11. (11. (11. (8x. (5x. (7x. (7x. (7x. (11. (16. (11
(21. (2x. (12. (44. (46. (61. (61. (56. (37. (23. (13. (30. (65. (70 (64. (44. (23. (14. (10. (8x. (25. (26. (28. (12. (14. (21. (24. (30. (18. (8x. (13. (15. (7x. (22. (21. (21

Figure 3: A nod automatically annotated with varying numbers of movement types in the 4th, 5th and 6th track from the bottom. 4th: only M=movement; 5th: Vertical, Horizontal or Other (V=Nod, UpNod, HeadBackward or HeadForward); 6th: all nine movement types (J=UpNod; N=Nod; HB=HeadBackward).

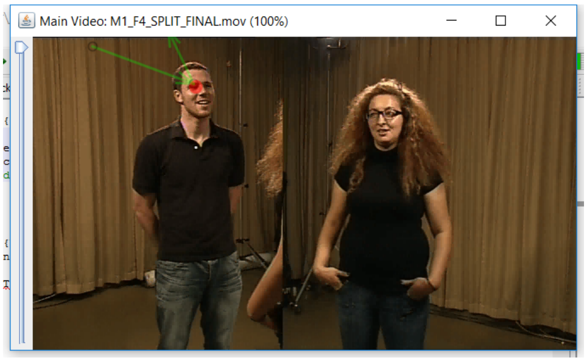


Figure 4: A single frame during a nod. The upward arrow indicates the direction and strength of the actual jerk. (The other arrow is an artefact caused by additional point data in the current jerk annotation, which is the highlighted box in Fig. 3)

for some head movement even better) was the classifier that differentiates between horizontal, vertical and ‘other’ head movements. The classifier used to predict all the nine classes reached the worst results. Average accuracies obtained by the three classifiers for different margin sizes (zero means no empty frames are added to the annotated movements, and 17 means that at most 17 empty frames were added) are shown in Table 2.

The automatic head movement annotations and manual annotations in general differ in granularity and assigned classes. Far from all the frames that are covered by a manual annotation also have automatic annotations. Sometimes a head movement is not detected at all, while in other cases much more than half of the frames covered by a manual annotation are also automatically annotated, see last column of Table 1. As explained above, bridging the annotations can help boost the classification results, but even without bridging the automatic annotations can give a good indication of the onset of a head movement, and also of the character of the movement.

In conclusion, we have proposed an approach to automatic head movement recognition in which jerk, the third derivative with respect to time, was added to the computation of velocity and acceleration in such a way that every single frame in a video is annotated with values for all three physical quantities.

These frame-wise data were combined with human annotations of head movement to create training and test sets for the development of classifiers aimed at identifying and labelling different types of head movement. We believe this approach is promising because i. it is based on a larger number of physiological measurements than previous attempts, ii. it does not use manually set, and therefore arbitrary, thresholds to establish movement onset or offset but learns such thresholds automatically from the data.

References

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

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

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, Patrizia Paggio, and Costanza Navarretta. 2016. Classifying head movements in video-recorded conversations based on movement velocity, acceleration and jerk. Submitted to COLING 2016.

Bart Jongejan, 2012. *Automatic annotation of head velocity and acceleration in Anvil*, pages 201–208. European language resources distribution agency, 5.

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).

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

L. Nguyen, J-M Odobez, and D. Gatica-Perez. 2012. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of ICMI’12*, Santa Monica, California.

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

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

Living with a Robot Pet: Context-Based Interpretation of Social Touch in Human-Robot Interaction

Merel M. Jung, Mannes Poel, Dirk K. J. Heylen
University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands
{m.m.jung, m.poel, d.k.j.heylen}@utwente.nl

Motivation

Touch plays an important role in interpersonal interaction touch were the modality can used to communicate emotions and other social messages [1, 2]. Extending social touch interaction to include interaction between humans and robots can result in more natural interaction. To behave socially intelligent, robots should not only be able to sense and recognize different touch gestures (e.g. [3, 5]) but should also be able to interpret touch to respond in a socially appropriate manner.

To interpret the meaning of touch in social interactions context is very important [2]. The complexity of the human tactile system allows for the same touch gesture to have different meanings as touch can also vary in its intensity, velocity, abruptness, temperature, location and duration [1]. Although there is no one-on-one mapping between touch gestures and their meaning, touch can have a clear meaning in a specific context [2]. Therefore, a touch gesture's meaning also dependents on factors such as the concurrent verbal and nonverbal behavior, the type of interpersonal relationship and the situation in which the touch takes place [2].

Robot seal Paro is probably the most famous example of a social robot that responds to touch [6]. Paro is equipped with touch sensors but only distinguishes between positive and negative touches [6]. However, research indicates that people use mostly positive forms of touch when interacting with another human [2] or a robot pet [7]. Furthermore, these positive forms of touch can have different meanings depending on the context, for example the intent of a touch could be affectionate, comforting/supportive or playful [2, 7].

In this paper we describe the setup and the planned analysis of a study in which participants interact with a robot pet companion in different emotional states. The aim of this study is to get more insight into the factors that are relevant to interpret human touch behavior within a specific social context.

Methods

Participants

In total 31 participants (11 female) volunteered to take part in the study. The age of the participants ranged from 22 to 64 years ($M = 34; SD = 13$) and all studied or worked at the University of Twente in the Netherlands.

Study setup

In the study participants were given four different scenarios in which they would come home in a particular emotional state, feeling either: depressed, stressed, relaxed or excited. These four emotional state were chosen as they span opposite ends of the valence and arousal scales: depressed (low valence,

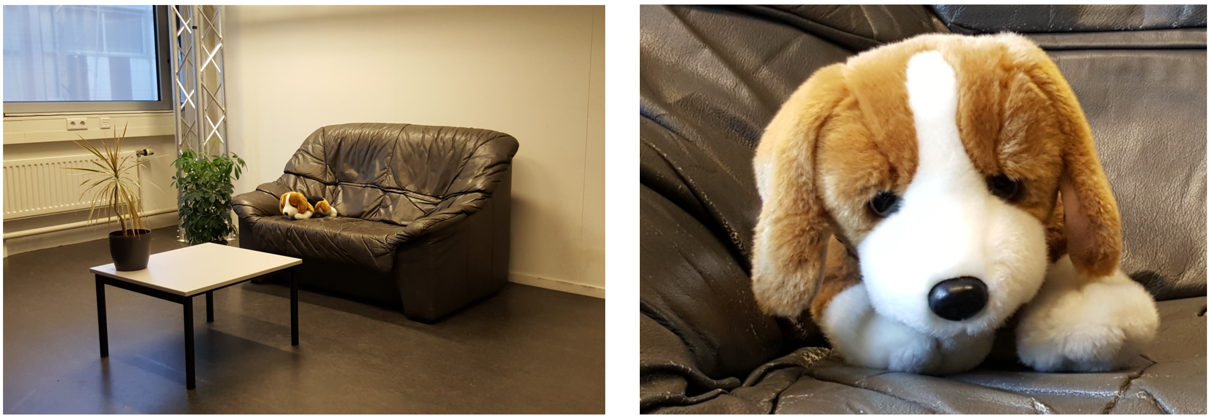


Figure 1: The living room setting with the robot pet on the couch (left) and the pet up-close (right).

low arousal), stressed (low valence, high arousal), relaxed (high valence, low arousal) and excited (high valence, high arousal) [4]. In each situation the participant was instructed to enter the 'living room' and sit down on the couch next to their robot pet and act out this situation as he or she sees fit. The living room setting consisted of a space of approximately 23 m^2 containing a small couch on which the robot pet was lying, a coffee table and two plants (see Figure 1). A stuffed animal dog (35 cm) was used as a proxy for a robot pet. Participants were instructed to focus on the initial interaction (a guideline of $\approx$ 30 seconds per situation was given) as the robot pet would not respond to them. The study had an within-subject design, instructions for acting out each of the four scenarios were given to each of the participants in random order. Interactions were recorded by two camcorders that were positioned in the corners facing the couch.

Afterwards, each of the four scenarios was played back to the participants and they were asked the following questions in an interview: (1) 'what message did you want to communicate to the robot?', (2) 'what response would you expect from the robot?' and (3) 'how could the robot express this?'. Participants also provided demographic information and filled in a questionnaire containing control questions about their ability to act out the scenarios. Furthermore, participants filled in another questionnaire in which they were asked to choose from a list of social messages which ones they had communicated to the robot pet in each of the four scenarios and which listed social messages they expected the robot pet to communicate to them in response. The six social messages (emphasize being together, greeting, play, seek/ give support, show affection and show appreciation) that could be selected were based on the categories of meanings of touch in interpersonal touch found in [2]. Lastly, participants were asked to imagine that they would get a robot pet as a gift that could react to touch and verbal commands. They filled in a questionnaire about the expected role that the robot pet would play in their life. Per participant the entire procedure took approximately 20 minutes.

Planned analysis

Video observations

Based on the video recordings of the interactions the general level of interaction with the robot pet in each of the scenarios will be assessed. Participants might have talked to the robot and touched it (i.e. a high level of interaction) or they might have either talked to or touched the robot or they could have chosen to completely ignore the robot pet. Because there was no set time limit for each of the interactions there might also be a difference in how long the participants interacted with the robot pet based on the situation. Although the focus of this study is on touch behavior, we will also look into other modalities as these can provide context for touch interpretation. Participants' behavior will be categorized to assess the use of multimodal signals in the different scenarios. The use of different types of touch behavior will be assessed such as: stroking the back of the robot pet, scratching the robot pet's belly and shaking the robot pet. Related to this is the body contact with the robot pet such as touching only with the hands while sitting next to it, putting the robot pet on his/her lap or holding the robot pet against the upper body. Also, the topics of speech can be of interest such as greeting, seeking interaction with the robot pet or talking about own emotional state. Furthermore, gaze behavior will be assessed such as looking at the robot pet while touching and establishing eye contact. Lastly, we will look at body expressions such as gesturing and posture.

Self reports

The interviews with the participants after watching back their interactions can give insight into the intent behind their actions. Social messages mentioned during the interview will be categorized and linked to the ones that are reported in the questionnaire. Participants might also have mentioned additional social messages that were not listed in the questionnaire. Furthermore, a behavioral model for the robot pet will be generated by matching the communicated social messages to the expected high-level response from the robot pet (i.e. the social message) and the behavioral response. Lastly, the expectations on the role that the robot pet would play in the life of participants can provide information about the types of interactions that a robot pet should be designed for.

Acknowledgments

This publication was supported by the Dutch national program COMMIT.

References

[1] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner. The communication of emotion via touch. *Emotion*, 9(4):566–573, 2009.

[2] S. E. Jones and A. E. Yarbrough. A naturalistic study of the meanings of touch. *Communications Monographs*, 52(1):19–56, 1985.

[3] M. M. Jung, X. L. Cang, M. Poel, and K. E. MacLean. Touch challenge '15: Recognizing social touch gestures. In *Proceedings of the International Conference on Multimodal Interaction (ICMI), (Seattle, WA)*, pages 387–390, 2015.

[4] J. A. Russell, A. Weiss, and G. A. Mendelsohn. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology*, 57(3):493–502, 1989.

[5] D. Silvera-Tawil, D. Rye, and M. Velonaki. Interpretation of the modality of touch on an artificial arm covered with an eit-based sensitive skin. *Robotics Research*, 31(13):1627–1641, 2012.

[6] K. Wada and T. Shibata. Living with seal robots – its sociopsychological and physiological influences on the elderly at a care house. *Transactions on Robotics*, 23(5):972–980, 2007.

[7] S. Yohanan and K. E. MacLean. The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. *International Journal of Social Robotics*, 4(2):163–180, 2012.

Grigory Kreydlin
Lidia Khesed
Russian State University for the Humanities, Moscow

The human body in multimodal communication: the semiotic conceptualization of hair

This paper explores several aspects of multimodal communication[1]. Primarily we study the correlation between natural language and its corresponding nonverbal cues, and the interplay of human systems (in the terminology of Apresyan 1995), particularly the physical, emotional and mental. The basic investigative method we use is the construction of a semiotic conceptualization of the somatic object «hair» in the Russian language, and in the language of Russian gestures.

The semiotic conceptualization of corporal objects and corporeality is a formal model that describes what common-language users think and say about their bodies and how they use their body parts and other somatic objects for various means of communication.

There are several reasons why hair was selected as the subject of our investigation. The Russian word *vólosy* is familiar to all Russians and is assimilated into both the Russian language and the culture. Hair, along with skin and nails, belongs to the class of so-called body covers, which have a number of specific features. The size, shape, color and functions of hair tend to reflect cultural beliefs about Russians and their typical communicative behavior. Hair is also involved in a number of Russian manual gestures and sign movements that vary from everyday corporeal practices and etiquette rituals to specific subclasses of gestures, such as grooming gestures, caressing gestures, gestures that express aggressive behavior, gestures that codify the emotional and mental states of people, etc.

The word *vólosy* and its derivatives are broadly presented in idioms, particularly in specific phraseological somatic units, or somatisms. The latter form the set of expressions with the name of a somatic object or one of its features (in the terms of Kozerenko, Kreydlin 1999, 2011). The collocations *ni ná volos* (‘not a bit of <…>’), *pokrasnét’ do kornéy volós* (‘blush to the hair roots’), *volosók k voloskú* (‘not a hair out of place’) not only convey meaningful characteristics of the hair, they also hand over information about the communicative behavior of Russians in certain situations.

[1] The research is conducted within the framework of the «Corporeal manifestations of mental and emotional activity of the human being» project supported by grant №16-34-00023 of the Russian Academic Foundation for the Humanities.

For describing the semiotic conceptualization of the human body and its parts, including hair, we have developed a special feature-based approach. It implies the identification and detailed analysis of three classes of body features. This includes physical (e.g. size, shape, colour), structural (e.g. texture, configuration, patronymic relations) and functional features (i.e. functions proper and dysfunctions) that manifest themselves in different language units and gestures. The feature-based approach correlates with the key aspects of multimodality as it reveals the essential links between Russian and its corresponding body code. The analysis of Russian gestures with hair is based on the Dictionary of Russian Gestures (DRG 2001), its ideology, methodology and general structure.

Our research follows the tradition of both the Moscow Semantic School (Y. D. Apresyan, I. A. Mel'čuk and others) and the ideas presented in the works on nonverbal semiotics (Kendon 1990, 2004; Kreydlin 2002, 2005; Müller, Cienki 2008).

The practical part of our research is based mainly on the analysis of the corporal data taken from the National Corpora of Russian Language (www.ruscorpora.ru). It also considers data from Russian explanatory and etymology dictionaries, idiom books and gestionaries (in the terminology of Poggi 2001) such as DRG 2001; Akishina, Kano 2010. Though the results of our research are relevant, first of all, for the Russian language and culture, most of them are equally relevant for other European languages. Thus, red hair is culturally marked in English, French and German; the gesture **to pull one's hair out <in disdain>** (Russian '*rvát' na sebé vólosy <v otcháyanii>*') is typical for most cultures. Examples of unique Russian language units related to hair are the names for children's hair – *volósiki* (it reflects its tenderness and softness) and a proverb *Rússkaya krasá – dlínnaya kosá* (lit. 'a Russian beauty is a long braid') applied to young women in Russian folklore.

Some physical and functional features of hair, along with their values, are described in detail as they are semantically and culturally marked. Among them are size, colour, shape, aesthetic and emotional functions. Thus, grey hair (*sedýe vólosy*) is prominent within the Russian culture: it is associated not only with a person's age, but also with his or her life experience and deep knowledge of people, e.g. the Russian expressions *dozhít' do sedýh volós* ('live until grey hair') and *uvazhénie k sedínam* ('respect for grey hairs'). The shape of hair can reflect not only some Russian aesthetic stereotypes, but also underlines the actual emotion of the person, e.g. *vsklokóchennye vólosy* ('disordered hair'). As for the aesthetic function of hair, it manifests itself in the rich culture of corporeal practices with hair: hairdo, hair care and decoration. The emotional function of hair is reflected, for example, in the gesture **to pull one's hair out**, which is usually performed in a situation of disdain and deep sorrow.

As for hair dysfunctions, an abnormal state of hair is an explicit marker of health or inner disease; in contrast, hair in a good state is a sign of human vitality.

A separate part of this paper is devoted to describing certain Russian gestures with hair. It deals mainly with caressing gestures, such as **stretching one's hair**, **rumpling one's hair**, and aggressive gestures, such as **dragging by the hair**, **pulling one's hair**. We pursue two aims in doing this. First, we describe the meanings and usage of gestures; second, we show the links between those gestures and certain Russian language units.

References

Akishina, Kano 2010 – Akishina A. A., Kano H. Zhesty i mimika v russkoy rechi. Lingvo-stranovedcheskiy solver' [Gestures and facial expressions in Russian spoken language. Cross-cultural dictionary]. – M.: URSS, 2010. 152 – P.;

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] in: Izbrannye trudy [Selected works]. Vol. II. Integral'noe opisanie yazyka i sistemnaya leksikografiya [Integral Description of Language and System Lexicography]. – M.: Yazyki russkoy kul'tury. – P. 348–388;

DRG 2001 – Slovar' yazyka russkih zhestov [Dictionary of Russian Gestures] / ed. Grigoriev N. V., Grigorieva S. A., Kreydlin G. E. – M.: Yazyki slavyanskoy kul'tury, Wien: Wiener Slavistischer Almanach, 2001 256 – P.;

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, 2004. 400 – P.;

Kreydlin 2002 – Kreydlin G. E. Neverbal'naya semiotica. Yazyk tela i estestvennyj yazyk [Non-verbal semiotics. Body language and natural language]. – M.: NLO, 2002. 581 – P.;

Kreydlin 2005 – Kreydlin G. E. Muzhchiny i zhenshchiny v neverbalnoy kommunikatsii [Men and women in nonverbal communication]. – M.: Yazyki slavyanskoy kul'tury, 2005. 224 – P.;

Kozerenko, Kreydlin 1999 – Kozerenko A. D., Kreydlin G. E. Russkie zhesty i russkie frazeologismy II (telo kak objekt prirody i telo kak objekt kul'tury) [Russian gestures and Russian idioms II (body as an object of nature and as an object of culture)] in: Frazeologiya v kontekste kul'tury [Phrazeology in cultural context] / Ed. V. N. Telia. – M.: Yazyki russkoj kul'tury, 1999. – P. 269 – 277;

Kozerenko, Kreydlin 2011 – Kozerenko A. D., Kreydlin G. E. Frazeologicheskiye somatizmy i semioticheskaya kontseptualizatsiya tela [Phraseological somatisms and semiotic conceptualization of the body] in: Voprosy yazykoznaniya Vol. 6, 2011. – M.: Nauka. – P. 54 – 66;

Müller, Cienki 2008 – Metaphor and Gesture. Cienki, A., Müller, C. eds. Amsterdam: John Benjamins, 2008. 307 – P.;

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. http://www.semioticon.com/virtuals/index.html, 2001 – 2002.

DOES L2 SPEECH GENERATE A HIGHER GESTURE RATE?

A STUDY OF DUTCH SPEAKERS OF ENGLISH

Varduhi Nanyan

Ghent University, Belgium

Varduhi.Nanyan@UGent.be

The study focuses on hand and arm gestures used by native Dutch (Belgian) speakers while narrating a cartoon to an interlocutor in their L1 and L2 English. Previous studies have addressed the co-speech gestures in English and Dutch with respect to the motion verbs (Stam 1999, Kellerman and Van Hoof 2003). In this study, however, our focus lies on identifying the differences in terms of the frequency of gesturing in L1 Dutch and L2 English and the effect L2 gestures might have on memory.

Recent research shows that gestures play a relevant role in L2 and bilingualism studies and give us some insight into “how communicative and psycholinguistic factors interact to shape the L2/bilingual system both as product and as process" (Gullberg 2010: 86). Gestures help us understand the role of input processing, attention and cognitive load in language acquisition and production processes (Gullberg 2010). Studies also evidence that people use gestures to facilitate the lexical retrieval process (Rauscher and Krauss 1996, Krauss and Hadar 1999) and to ease the cognitive load on verbal working memory (Gillespie et al. 2014).

L2 speech tends to be more hesitant containing longer pauses and tongue slips which reduce the working memory capacity (Payne et al. 2012). Reduced memory capacity implies a higher frequency of gesturing to lighten the load on the memory and ease the efforts of language processing (Gillespie et al. 2014). This allows us to assume that L2 proficiency might have an effect on the working memory by putting less cognitive load on it.

In this study our aim is twofold. Given the suggestions above it seems plausible to assume that bilinguals will use more gestures in their L2 than in L1 speech to facilitate the cognitive load during the second language production. To test this hypothesis, we compare the frequency of the gestures used by the speakers in their L1 storytelling and that of the same speakers in their L2 speech by looking at the overall number of gestures as well as at the numbers of the specific categories of gestures (iconic, metaphoric, deictic, beats). Furthermore, basing on the suggestions that proficiency can mediate lexical access in speech production (Prebianca 2014) and that proficiency can influence “the way speakers gesture in relation to speech” (So, Kita, Goldin-Meadow 2013: 538) as well as taking into account that gestures lighten the load on the memory (Gillespie at al. 2014), we test whether proficiency has a bearing on the frequency of gesturing in L2 use.

To elicit gesture an experiment was designed during which seventeen informants (all Ghent University students, aged 20-26) were asked to watch a short cartoon clip and then retell it in two languages - first in English, then in Dutch - to a listener who was naïve. The informants told the stories in both languages to the same listener who was neither a native speaker of Dutch nor of English. They were not provided with any information about the ethnic and linguistic background of the listener. The narrations were videotaped for coding. We used ELAN for annotating the video data. The data were transcribed focusing on the verbal utterances and the co-occurring gestures.

The English language proficiency of the participants was determined through self-reported and behavioural measures. The self-reported data have been gathered through the Language Experience and Proficiency Questionnaire (LEAP-Q) elaborated by Blumenfeld and Kaushanskaya (2007). The test focuses on the language acquisition, competence, experience and use. The behavioural data have been collected through two standardized tests - the Quick Placement Test (QPT) and LexTALE. QPT (Athanasopoulos 2007) measures the reading skills, vocabulary, and grammar. LexTALE (Lemhöfer and Broersma 2012) assesses the comprehension and the vocabulary knowledge of the participants. For measuring the language production with respect to the lexical access (Levelt 1989, Dell et al. 1999) a Multilingual Naming Test (MINT) (Gollan et al. 2012) designed by us was used. The participants were asked to complete the four proficiency tests following the video part. The tests were presented in the same order to all the participants.

The results of the proficiency tests were consistent in terms of the L2 proficiency level with a strong correlation (r=.72) between MINT and LexTALE. The results of these two tests were also predictive of the QPT results. Taking this into account, we have considered the QPT results as an estimated proficiency level for each participant. Another reason why it is convenient to use these results is because they are presented within the Common European Framework of Reference for Languages (CEFR), which is influential in Europe and beyond (Hulstijn 2011). We have divided the participants into two main proficiency groups: intermediate and advanced, also taking into account the CEFR levels.

To keep the results comparable the rate of gesturing was determined following Kita's (1993) and Gullberg's (1998) model by counting the number of gestures per clause.

We used a paired-sample t-test (between subjects) to compare the difference between the gesture rate in English and in Dutch. We applied this approach to control for the variance between individuals. We did this for the overall gesture rate first, and then looked at the specific categories.

The results reveal that Dutch speakers tend to use more gestures in their L2 English speech (p = .058, r=.78). Specifically, we find significant differences in the categories of the iconic (p = .002) and deictic gestures (p = .048) whereas the differences found in the categories of the metaphoric gestures (p = .58) and beats (p = .64) are marginal. Our findings are quite different from Marcos' (1979) results. This study reported that Spanish-English and English-Spanish speakers deployed more gestures of all types in their L2, but that significant differences were found for beats. One reason to account for the cross linguistic differences among the languages in terms of the gesture rate in L2 production would be to assume that they are culturally and typologically determined (Kendon 1981, 1992, Efron 1972).

Further analysis suggests that the different measures of L2 proficiency are strongly inter-correlated, but there are no significant differences between the proficient L2 speakers and their less fluent peers in terms of the gesture rate in L2 (p >.05). The results might be due to the tendency of the less proficient L2 speakers to keep their narrations shorter while the narrations of the more fluent informants tended to be longer and more detailed.

The findings provide at least partial support for the Verbal Working Memory (VWM) and the Lexical Retrieval theories. The study further supports Kita's (1993) claim that bilingual (multilingual) speakers tend to use more gestures in L2 speech. One important factor that might have triggered the difference between the L1-L2 gesture rate could have been determined by the fact that the informants were asked to narrate the story first in English, then in Dutch, hence, they might have been more at ease when narrating the cartoon in Dutch, due to which the speakers have produced fewer gestures in L1. Given *the lexical retrieval* and the *cognitive load facilitation hypotheses* first narrations (see also Brown and Gullberg 2013) and the common ground (Holler and Wilkens 2011) might have affected the increased gesture rate in L2.

Basing on McNeill's (1992) proposal that gesture and speech are tightly connected as they co-occur as a result of the interaction of two mental operations - a linguistic and an imagistic one - and that gesture and speech are realized by shared processing mechanisms, we can suggest that "gesture production is linked to some of the same VWM-related processing resources recruited during message formulation and grammatical planning in language production" (Gillespie et al. 2014: 178). In this paper we do not explain how this interaction takes place but adhering to Krauss et al. (1999, 2000) assume that gesture facilitates lexical access and aids in activating the spatial representations and the working memory.

This work sheds light on the cross-linguistic differences in terms of the gesture rate in L2 speech and could serve as a tiny window into a bilingual's cognitive abilities. For further research it would be interesting to conduct an experiment with a counterbalanced measures design whereby the subjects would be provided other stimuli (i.e. another cartoon clip) and asked to narrate it in a reverse order: first in L1 Dutch, then in L2 English. This would give one a clue whether the gesture rate in L1 and L2 speech is dependent on the order in which the stimuli are presented. To account for the cross-linguistic differences and to answer the research question in a more comprehensive way it would also be interesting to replicate the study with English L1 speakers speaking Dutch (Belgian) as their L2. It should also be noted that the gesture rate patterns in L1 and L2 as well as the effect of proficiency on L2 gesturing might vary cross-linguistically and studying a combination of other languages might bear quite different results. This can be avenue for future research, too.

References

Athanasopoulos, P., 2006. Effects of the grammatical representation of number on cognition in bilinguals. *Bilingualism: Language and Cognition*, 9, 89–96.

Blumenfeld, M., Kaushanskaya, 2007. Language Experience and Proficiency Questionnaire (LEAP-Q), *The Journal of Speech Language and Hearing Research*, 50(4), 940-967

Brown, A., & Gullberg, M. 2013. L1–L2 convergence in clausal packaging in Japanese and English. Bilingualism: Language and Cognition, 16, 477-494. doi:10.1017/S1366728912000491

Dell, G. S. et al., 1999. Connectionist Models of Language Production: Lexical Access and Grammatical Encoding, *Cognitive Science*, Vol 23 (4), 517–542

Gillespie, M., James, A.N., Federmeier, K. D., Watson, D. G., 2014. Verbal working memory predicts co-speech gesture: Evidence from individual differences, *Cognition 132(2),* 174–180

Gollan, T. H., Weissberger, G., Runnqvist, E., Montoya, R. I., & Cera, C. M., 2012. Self-ratings of spoken language dominance: A multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition*, 15, 594–615.

Gullberg, M. 1998: Gesture as a communication strategy in second language discourse: a study of learners of French and Swedish. Lund: Lund University Press.

Gullberg, M. 2010. Methodological reflections on gesture analysis in SLA and bilingualism research. *Second Language Research,* 26(1), 75-102

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.

Hulstijn, J. H. 2011. Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. Language Assessment Quarterly, 8, 229–249

Kellerman, E. & Van Hoof, A.-M., 2003. Manual accents. *International Review of Applied*

Linguistics 41, 251-269.

Kita, S. 1993. Language and thought interface: A study of spontaneous gestures and Japanese mimetics, University of Chicago, Department of Psychology, Department of Linguistics

Krauss, R.M., Hadar, U., 1999. The role of speech-related arm/hand gestures in word retrieval. In, R. Campbell & L. Messing (Eds.), *Gesture, speech, and sign*, Oxford: Oxford University Press; 93-116.

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

Lemhöfer, K. & Broersma, M., 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44(2), 325-343.

Levelt, W. J. M. 1989. Speaking: From intention to articulation. Cambridge, MA: MIT Press

Marcos, L. R., 1979. Nonverbal behavior and thought processing. Archives of General Psychiatry, 36 (9), 940–943.

Payne, J.S., Whitney, P. J., 2002. Developing L2 Oral Proficiency through Synchronous CMC: Output, Working Memory, and Interlanguage Development, *CALICO Journal*, 20 (1), 7-32.

Pika, S., Nicoladis, E., Marentette, P. F., 2006. A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer? *Bilingualism: Language and Cognition*, 9 (3), 319–327.

Prebianca, G.V.V., 2014. Exploring the relationship between lexical access and proficiency level in L2 speech production, *Trab. linguist. apl*. vol.53 no.2, Campinas

Rauscher, F.H., Krauss, R.M., Chen, Y., 1996. Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science* 7, 226–231.

So, W.C., Kita, S., Goldin-Meadow, S., 2013. When do speakers point? The role of language proficiency in semantic relation between pointing gestures and speech. *Journal of Psycholinguistic Research*, 42, 6, 581-594.

Stam, G., 1999. Speech and gesture: What changes first in L2 acquisition? Paper presented at the Second Language Research Forum, Minneapolis, MN.

Speech pauses, gestures and audience laughter in English humorous speech

Costanza Navarretta

University of Copenhagen
costanza@hum.ku.dk

1. Introduction

This study is about the use of speech pauses and gestures in humorous speech as means to engage the audience and present the message in an effective way. The data which we investigate are two speeches by the American president Barack Obama at the Annual White House Correspondents' Association Dinner. According to tradition, the president gives a speech or sends a video where he mocks himself, his collaborators, adversaries, and the press corps. Audience engagement is here measured in terms of the immediate audience response to a message via laughter and/or applause.

The speech pauses which we include in the study are silent pauses and filled pauses such as *um*, *ah*, and *uh*. We also account for various gesture types (head movements, facial expressions and hand gestures), and we analyze their relation to speech pauses.

Pauses and gestures have multiple and often co-occurring functions both in speech production and speech perception. Gestures have a central role in interaction management [1], and they contribute to both the content and structure of discourse [2, 3]. Pauses are voluntary or involuntary signals regulating the interaction [4, 5], and they can signal that speakers are planning and structuring their message [6, 7], or are presenting difficult concepts [8]. In a study of intonation and silent pauses, Hirschberg and Nakatani [9] find that pauses are often used as markers of discourse structure in both read and spontaneous English speech. The role of ungrammatical silent pauses and rate of articulation in the sitcom *Friends* is investigated by inter alia [10, 11] who conclude that sitcoms exhibit features of both spoken and written discourse since they are based on written texts but they are acted as spontaneous speech. Finally, pauses and their timing in comedy have been addressed by both researchers and comedians, inter alia [12, 13] who point out that pauses are not only means to structure and emphasize the discourse, but they also give the audience time to reflect on and appreciate the conveyed message.

2. Data

The two speeches by Barack Obama at the White House Correspondents' Association Annual Dinner were held in 2011 and 2016, respectively. We use the videos produced by the White House which are available at `http:\\www.WH.gov`, and record the president frontally as shown in figure 1. We converted the recordings to avi format and extracted wav audio files. Silent pauses were automatically transcribed using a PRAAT in-built script [14]. We found that the best silent threshold for delimiting silent pauses in these data was -35.0 dB with the minimum silent interval duration set at 0.2 seconds. Then, the speeches were transcribed replacing Obama's words and the audience's reaction into the appropriate sound parts of the automatically produced TextGrid file. The automatic annotation of silent pauses was also corrected. The resulting transcriptions consist of speech segments (one or more speech tokens), silent or filled pauses, and audience response. Successively, we annotated Obama's gestures in the ANVIL tool [15]. The duration of the annotated 2011 video segment is 13 minutes and 22 seconds, while the duration of the annotated 2016 video segment is 30 minutes.

Figure 1: A snapshots from the 2016 speech

The shape and semiotic type of gestures were annotated following the MUMIN annotation scheme [16]. Table 1 shows the shape features. The semiotic types annotated are *indexical deictic*, *indexical non-deictic*, *iconic*, *iconic metaphoric* and *symbolic* and are inspired by [17].

3. Preliminary analysis

In table 2 the frequency of speech tokens, pauses and gestures are given as total occurrences and occurrences per second. When calculating the ratio speech token per sec-

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
TrajeLeftHand	LHForward,LHBackward,LHSide ,LHUp,LHDown,LHComplex,LHOther
TrajeRightHand	RHForward,RHBackward,RHSide RHUp,RHDown,RHComplex,RHOther
PalmOrientation	PalmUp,PalmDown,PalmSide, PalmVertical,PalmOther

ond, we did not consider the time in which the audience laughs and/or applauds. The resulting speech duration is 8 minutes for the 2011 data and 19 minutes for the 2016 data. Many of the head movements co-occur with movements of the upper body, which have not been annotated. Not surprisingly, the speech rate (words and fillers) is nearly the same in the two speeches. In the 2011 data, however, silent and filled pauses are more frequent than in the the 2016 data. That is Obama uses more frequently silent pauses in the 2011 speech than in the 2016 speech, but they have shorter duration in the former speech than in the latter one. The low frequency of fillers in these speeches was expected because Obama is reading from a manuscript. The analysis of their occurrences shows that Obama uses them consciously to emphasize what he has said or what he is going to say. In a couple of cases he repeats the same filler. The temporal ratio of the audience's 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 2011 and 40% of the speech duration in 2016.

Table 2: Frequency

token	11 #	11 #/s	16 #	16 #/s	Tot	Tot/s
speech	1059	2.21	2531	2.22	3590	2.33
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	50	0.1	84	0.007	134	0.08
hand	51	0.11	237	0.21	289	0.18

Obama moves the head continuously turning it to the right or to the left to address the whole audience in the room. Sometimes he turns his body to address the people sitting behind him. During speech pauses, or while the audience is laughing, Obama often moves his head forward and looks at his manuscript. The frequency per second of head movements in the two speeches is the same, while Obama produces more hand gestures and less facial expressions in the 2016 than in the 2011 speech.

The most common facial expressions in these data are smiles and expressions in which Obama retracts his lips while listening at the audience's applauses and/or laughs. Obama laughs only few times. All kinds of hand gestures are produced, the must frequent gestures being deictics, iconics, metaphorics, and beats. Obama also produces gestures discourse structuring gestures.

4. Discussion

The preliminary analysis of the multimodally annotated Obama's speeches from 2011 and 2016 indicates that Obama produces more silent pauses in 2011 than in the 2016, while he receives the same amount of feedback by the audience in terms of the temporal ratio of his speech and of the audience reaction in the form of applause, laughter and cheers. Thus indicates that the frequency of silent pauses is not related to the frequency of audience response.

A first analysis of the silent pauses in the two speeches indicates that most of these pauses delimit grammatical phrases (nominal, adjectival, verbal, adverbial and clausal phrases). A number of pauses delimit topic shifts confirming thestudy by [9]. Furthermore, we found a large number of pauses that precede 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 [18]. Moreover, these pauses are not used to fake spontaneous speech as in sitcoms [11]. Instead, they are so called *emphatic pauses* which emphasize the following speech segment. Finally, a number of pauses that follow a word or a phrase are used to let the audience get the point, and in most cases they are followed by the audience's laughter and/or applause. Filled pauses are few and Obama uses them consciously to emphasize what he has just said. Finally, a number of short pauses follow the audience response, indicating that Obama adapts his speech to the audience's reaction.

The multimodal analysis of the speeches indicates that communicative gestures co-occurring with speech pauses are seldom, while Obama often uses speech pauses to look at his manuscript. Therefore, silent pauses also co-occur with head forward movements. Thus, these data confirm research by inter alia [19] who find that in English and Italian different data types speech pauses are often co-occurring with speech holds. They propose that speech pauses and gestural holds have parallel functions of introducing new information.

Obama interacts actively with his audience, talking and pointing to individuals in the room. In most cases, he does not laugh with the audience. Most exceptions in these data are jokes involving Obama's wife, Donald Trump and other presidential candidates.

Most of the gestures produced by Obama, and especially his hand gestures in the 2016 speech, are those typical of speeches, with the exception of deictics pointing to people in the room. Obama is often serious while presenting his jokes, and this seriousness seems to be one of the means to make the audience laugh. However, the multimodal behavior of Obama in these speeches should be compared with his behavior in more traditional speeches in order to assess whether this is the case.

Since many pauses are used to emphasize the coming speech, we have tested whether is possible to predict audience laughter from Obama's speech segments, pauses and gestures. The results of our first experiments in which we train a Naive Bayes and a Logistic algorithm on the transcripted data (trigrams of speech segments, pauses and audience response) and tested on data in which the audience response is not annotated , show that information on the alternation of speech pauses can be used to some extent to predict the audience reaction (laughter). More details of these experiments will be presented in the final version of the paper.

Finally, it must be noted that we have not analyzed the speech content and we have not included intonation features which are central aspects when talking about humorous speech. The type of audience is also relevant with respect to their reaction to the jokes. These features should also be included in future investigations of the data.

5. References

[1] J. Allwood, "The Structure of Dialog," in *Structure of Multimodal Dialog II*, M. M. Taylor, F. Neél, and D. G. Bouwhuis, Eds. Amsterdam: John Benjamins, 1988, pp. 3–24.

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

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

[4] S. Duncan and D. Fiske, *Face-to-face interaction*. Hillsdale, NJ: Erlbaum, 1977.

[5] H. H. Clark and J. E. Fox-Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–11, 2002.

[6] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous english speech," *Word*, vol. 15, pp. 19–44, 1959.

[7] W. Chafe, "Cognitive Constraint on Information Flow," in *Coherence and Grounding in Discourse*, R. R. Tomlin, Ed. Amsterdam: John Benjamins, 1987, pp. 20–51.

[8] S. R. Rochester, "The significance of pauses in spontaneous speech," *Journal of Psycholinguistic Research*, vol. 2, pp. 51–81, 1973.

[9] J. Hirschberg and C. Nakatani, "Acoustic Indicators of Topic Segmentation," in *Proceedings of ICSLP-98*, Sidney, 1998.

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

[11] M. Bilá, "A Comparative Analysis of Silent Pauses and Rate of Articulation in the Discourse of Sitcom," *Discourse and Interaction*, 2014.

[12] J. Sankey, *Zen and the Art of Stand-Up Comedy*. New York: Routledge, 1998.

[13] B. Oliver, *The Tao of Comedy: Embrace the Pause*. Oliver, 2013.

[14] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, 2009, retrieved May 1, 2009, from http://www.praat.org/.

[15] M. Kipp, "Gesture generation by imitation - from human behavior to computer character animation," Ph.D. dissertation, Saarland University, Saarbruecken, Germany, Boca Raton, Florida, dissertation.com, 2004.

[16] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "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*, vol. 41, no. 3–4, pp. 273–287, 2007.

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

[18] R. Krauss, Y. Chen, and R. F. Gottesman, "Lexical gestures and lexical access: a process model," in *Language and gesture*, D. McNeill, Ed. Cambridge University Press, 2000, pp. 261–283.

[19] A. Esposito and A. M. Esposito, "On Speech and Gesture Synchrony," in *Communication and Enactment - The Processing Issues*, ser. LNCS, A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, Eds. Springer-Verlag, 2011, vol. 6800, pp. 252–272.

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

Thomas Ousterhout

Center for Language Technology
University of Copenhagen
Njalsgade 140, 2300 Copenhagen S
Email: tko@hum.ku.dk

Semantically meaningful pictures, which are commonly called emoticons or emojis, are becoming increasingly popular for people to include in personal text messages. These emoticons, similarly to gestures in face-to-face communication, add bodily behavior to the text message as supplemental information which would not normally be there. Evolutionary psychology explains how these many types of bodily behaviors including hand gestures, body language, and facial expressions, have been existent long before the development of spoken language [1], [2]. While it is presumed that the inclusion of semantically meaningful pictures of bodily behavior in text is an attempt to replicate this multimodal communication.

A neurophysiological phenomenon occurs in relation to semantic priming, which can be measured through EEG and involves event-related potentials (ERPs). The EEG component called the N400 ERP is a negative going deflection with a posterior and broad scalp distribution that usually occurs between 200 and 600 ms post stimulus onset of a critical/target word and is affected by semantic relations [3], [4]. Just like facilitated reaction time with related pairs, there is a smallerN400 effect when related versus unrelated pairs are presented as well, meaning that the negative deflection is not as large [5], [6]. The more unrelated the probe word is, the larger the N400 deflection will be.

There have also been several semantic priming N400 studies investigating how visual gestures are processed in relation to verbal material [7], [8]. Due to the fact that symbolic gestures and words are reciprocal when they are semantically congruent [9], [10] used symbolic gestures as primes and found in their investigation that they perform just as well as verbal material in producing the N400 modulation. It has been shown that N400 effects can be created with probe words in relation to the semantic content to previous sentences. It has also been demonstrated that primes in this previously presented semantic content can be symbolic gestures. Thus, the hypothesis of this experiment expands on the idea in that a picture, similar to an emoticon, depicting a symbolic gesture or emotion, presented in the middle of a sentence, will function as the salient prime in the sentence and produce and N400 effect to subsequently presented semantically incongruent probes.

The participants for this study included 11 volunteers (6 females, 5 males), 10 were right handed, and they were aged between 26 and 57 years (mean = 36 years). The study used 21 priming sentences that each had a congruous and incongruous subsequent probe sentence. Each prime sentence was short with only 3 or 4 words and each had a picture of either an emblematic gesture, or a happy/sad face emoticon between the concluding salient word and the prior word. The probes consisted of a three word sentences, where one word was displayed at a time, and the final word would be directly congruent or incongruent with the picture in the prime sentence.

Using a paired t test, these reaction time means were statistically significantly different with P < .0006. Also there was a statistically significantly different mean amplitude in the N400 region where incongruent probes were more negative than congruent.

With these findings it seems evident that the presentation of semantically meaningful gestures can activate cognitive feature processing stored in working memory and thus provide facilitated cognitive recognition processing in subsequently presented words related to the objects with those features. The results of this present study are consistent with experiments that investigate high cloze sentences, which are those where there is a preferred ending which attenuates the N400 amplitude, versus low cloze sentences which have unlikely endings and thus produce large N400 amplitude deflections.

[1] G. Rizzolatti and M. A. Arbib, "Language within our grasp," Trends in neurosciences, vol. 21, no. 5, pp. 188–194, 1998.

[2] M. Tomasello, "Primate cognition: introduction to the issue," Cognitive Science, vol. 24, no. 3, pp. 351–361, 2000.

[3] M. Kutas and S. A. Hillyard, "Brain potentials during reading reflect word expectancy and semantic association," 1984.

[4] S. Bentin, G. McCarthy, and C. C. Wood, "Event-related potentials, lexical decision and semantic priming," Electroencephalography and clinical Neurophysiology, vol. 60, no. 4, pp. 343–355, 1985.

[5] J. Kounios and P. J. Holcomb, "Structure and process in semantic memory: evidence from event-related brain potentials and reaction times." Journal of Experimental Psychology: General, vol. 121, no. 4,p. 459, 1992.

[6] P. J. Holcomb and J. E. Anderson, "Cross-modal semantic priming: A time-course analysis using event-related brain potentials," Language and Cognitive Processes, vol. 8, no. 4, pp. 379–411, 1993.

[7] S. D. Kelly, C. Kravitz, and M. Hopkins, "Neural correlates of bimodal speech and gesture comprehension," Brain and language, vol. 89, no. 1,pp. 253–260, 2004.

[8] Y. C. Wu and S. Coulson, "Meaningful gestures: Electrophysiological indices of iconic gesture comprehension," Psychophysiology, vol. 42,no. 6, pp. 654–667, 2005.

[9] F. Barbieri, A. Buonocore, R. Dalla Volta, and M. Gentilucci, "How symbolic gestures and words interact with each other," Brain and Language, vol. 110, no. 1, pp. 1–11, 2009.

[10] P. Bernardis and M. Gentilucci, "Speech and gesture share the same communication system," Neuropsychologia, vol. 44, no. 2, pp. 178–190, 2006.

BCI effectiveness test through N400 replication study

Thomas Ousterhout

Center for Language Technology
University of Copenhagen
Njalsgade 140, 2300 Copenhagen

Brain-computer interfaces (BCI) are beginning to appreciate much popularity and interest in the research and consumer markets. There is however a large practical and functional variability between commercial grade systems and medical grade EEG headsets. This study was conceived primarily in order to validate the signal quality of commercially available and user friendly neuroimaging technology as a brain-computer interface (BCI) and electroencephalography (EEG) research tool in comparison to medical grade devices.

EEG is a cognitive measurement tool used to detect and measure the electrical signals in the brain when neurons communicate with each other. Invasive, cortically-implanted electrodes, allow for amore precise method of measuring brain activity, however noninvasive scalp electrodes allow for a much more appropriate scientific method for the average researcher and user [1].

In face-to-face communication, people typically simultaneously utilize both modes of auditory and visual information which is called multimodal communication. Examples of unimodal communication would be sign languages which is purely visual, and speaking over the phone which is purely auditory. The auditory modality typically provides the most information content in face-to-face communication and thus usually is classified as the dominant modality of communication. However, the visual cues are very important and sometimes necessary to understand fully what the intended message is [2].

Some aspects of semantic meaning can be measured with EEG by looking at something called event-related potentials (ERPs), which are EEG amplitude deflections in the brain produced in response to certain events or stimuli. The N400 ERP is a negative deflecting component occurring 400 ms after the onset of a auditory or visual stimulus that has been studied for over thirty ears to measure semantic processing [3], [4], [5], [6].

The present study attempts to reproduce the results found by [6] 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. The hypothesis is that since the N400is such a large ERP, even with the poor resolution of the 14-channel Emotiv, in comparison to 59-channel medical grade EEG scalp cap used by [6], the Emotiv will still be able to detect the N400 when comparing the meaningless hand postures with the meaningful ones.

The current study used 16 participants aged between 20-37 years (mean = 26.9), 9 were males, 7 were females and all were right handed. The participants were native English or fluent English speaking adults at the University of Copenhagen who study and/or speak English daily. The experimental stimuli was acquired courtesy of [6] since it was that study which was replicated. The stimuli consisted of photos 11 different meaningful and meaningless hand postures. The results show that there was an effect for meaningfulness and thus was identified as a N400.

This study investigated the effectiveness of using a BCI as an EEG acquisition tool which has fewer electrodes and thus poorer resolution, and tested in a more real life environment with noise. The ERP under investigation was the N400 effect regarding meaningless hand gestures compared to meaningful hand gestures made up of emblem, iconic, and deictic gestures. This study tested the effectiveness by replicating the paradigm of [6] and reproduced the results 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.

[1] B. C.-T. Lin, L.-W. Ko, J.-C. Chiou, J.-R. Duann, R.-S. Huang, S.-F. Liang, T.-W. Chiu, and T.-P. Jung, “Noninvasive neural prostheses using mobile and wireless eeg,” Proceedings of the IEEE, vol. 96, no. 7, pp.1167–1183, 2008.

[2] H. H. Clark, Using language. Cambridge university press Cambridge,1996, vol. 1996.

[3] M. Kutas and K. D. Federmeier, “Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp),” Annual review of psychology, vol. 62, p. 621, 2011.

[4] C. C. Duncan, R. J. Barry, J. F. Connolly, C. Fischer, P. T. Michie, R. N¨a¨at¨anen, J. Polich, I. Reinvang, and C. Van Petten, “Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, p300, and n400,” Clinical Neurophysiology, vol. 120, no. 11, pp. 1883–1908, 2009.

[5] M. Kutas and K. D. Federmeier, “Electrophysiology reveals semantic memory use in language comprehension,” Trends in cognitive sciences, vol. 4, no. 12, pp. 463–470, 2000.

[6] T. C. Gunter and P. Bach, “Communicating hands: Erps elicited by meaningful symbolic hand postures,” Neuroscience Letters, vol. 372,no. 1, pp. 52–56, 2004.

Forte, *piano*, *crescendo*, *diminuendo*: signals of intensification and attenuation in orchestra and choir conduction

Isabella Poggi
Dipartimento di Filosofia, Comunicazione e Spettacolo
Università Roma Tre - Roma

1. Attenuation and intensification in music conducting

The expressions of attenuation and intensification can be defined as the signals aimed at informing that the level of a certain quantity or intensity is either lower or higher than expected: this is the meaning, for instance, of expressions such as “less”, “not so”, or “strongly”, “highly”, or emphatic batonic gestures, oscillating or rounding vagueness gestures, which beside their literal meaning may be aimed at mitigation or other rhetorical functions. But intensity or quantity may not only be the object of informative speech acts or body communicative acts: they can also be the object of requests, for instance when a teacher asks pupils not to speak so loud, or when a stranger asks you to slow down your high speech rate. Another case of requestive attenuation and intensification signals are those performed in a peculiar communicative situation: the gestures and other body signals performed in conducting.

The Conductor’s language is an intriguing case of communicative interaction, as pointed by Ashley (2000), who applied Grice’s model to describe it; its multimodality has been studied by Boyes Braem and Braem (1999), who analysed the metaphorical import of the conductor’s gestures, and by Dineen (2011), concerning the economy of gestures, and Luck (2011), with his computational model of temporal gestures. Poggi (2011) stressed the conductor’s gestures, face, posture, gaze, in both concert and rehearsal, provide information about who should sing or play, when, what semantic content to express by words and music, what melody, rhythm, tempo, timbre, intensity, expression, musical structure to produce, and how.

Among all this information conveyed by conductors, a specific kind concerns sound intensity: this work analyses the signals for *piano, forte, crescendo, diminuendo*, and other dynamic indications provided by the conductor’s hands, gaze, head, body movements.

2. A shared lexicon

Previous works have argued that the signals used in musical performance are not totally idiosyncratic but systematic and shared. Concerning a Pianist’s body movements, Poggi (2006) showed that a “lexicon” can be outlined i.e., a list of systematic correspondences between signals and meanings. Some of a Pianist’s body movement are utterly communicative, and can be attributed a performative and a content; some express cognitive (e.g. concentration) or emotional states felt about playing (satisfaction, worry) or to be conveyed in music (sadness, mirth); others, accompanying the technical movement needed to produce a particular sound, tymber, rhythm, “help” the pianist to perform it better: e.g. *frowning*, an expression of anger, by evoking anger mobilizes energy, thus helping to play “forte”. The same movements of trunk, gaze, head and face recurrently have the same goal or convey the same meanings both across different passages in a music piece, and across concert and rehearsal. But if body signals are so systematic in the individual “lexicon” of a single musician, they must be more so in Conductors, whose signals do not have an expressive function only but a communicative one: if s/he must convey what, when, and how to sing or play, then his/her signals must be shared and form a common lexicon to allow an Italian Conductor be understood by a German or a Chinese orchestra.

3. The lexicon of attenuation and intensification in orchestra and choir conduction

Starting from the hypothesis that a lexicon of signals in music conduction can be outlined, within the Conductor’s body signals conveying melody, rhythm, tempo, timbre, intensity, expression, musical structure, we focused on musical intensity, in particular on the dynamic indications for *forte, piano, crescendo, diminuendo* and conducted an observational qualitative study to analyze the bodily signals of sound attenuation and intensification performed by orchestra and choir conductors.

3.1. Corpus, hypotheses and method

Ten fragments of orchestra and choir conduction by three different conductors were analyzed: two fragments taken from rehearsals by Riccardo Muti and Leonard Bernstein on YouTube, and eight (4 from concert, 4 from rehearsal) by Alessandro Anniballi, the conductor of an amateur choir. The fragments in total last 122'35", 27'55" of concert and 94'40" of rehearsal, respectively, and through a dedicated coding scheme each signal of intensity was analyzed in detail.

Beside time in the video and the dynamic indication of the musical score, the concomitant words sung are annotated, the modality under analysis, and a description of the signal in terms of its parameters: e.g., posture is described as "*Bust forward, shoulders closed, head forward downward*". Then, the "body meaning" is annotated, i.e., the goal of the body movement performed: here, the posture is that of someone curling onto oneself, thus making oneself smaller. Hence the meaning of the signal is interpreted: curling onto oneself to *get smaller* means "softer", i.e., "make a *smaller* sound". Finally the signal is classified in terms of its underlying semiotic device. For instance, making oneself smaller is an iconic gesture that exploits a "transmodal shift", from space to sound: a smaller body is similar to a smaller (softer) music; a body taking less room recalls a sound taking less energy.

Based on this analysis, the work tried to find out if the same signals or aspects of signals systematically convey the same meanings throughout music performance. For each of the four meanings *forte, piano, crescendo, diminuendo*, a hypothesis was put forward about the possible corresponding body signals by hands, face, gaze and posture. For example, a hypothesis was that *index finger upward over lips* – a symbolic gesture for "silence" – typically asks for *piano*; a *high level of muscular tension* generally asks for *forte*.

Each of the hypotheses put forward as to the meaning of conductors' signals was tested, on the one side, by careful analysis of videos without audio: the researcher first makes his hypothesis about a signal, and then verifies, with audio on, if the corresponding sound (and/or the dynamic indication in the score) confirms it. On the other side, in the score of the music played the Author's indications of *piano, forte, crescendo, diminuendo*, are searched, and for each indication what signals are used is checked on the corresponding video.

3.2. Results

This analysis resulted into a list of signals, performed by head, hands, gaze, face, posture and body movements, with their corresponding meanings, and into a first overview of the semiotic devices exploited in those signals.

From a quantitative point of view, in the fragments analyzed signals for intensification are more frequent than ones for attenuation: respectively, 62 signals for "forte" and 4 for "crescendo", 38 for "piano" and 5 for "diminuendo". Signals are not all of different "types"; sometimes, the same signal "type" has several tokens, several occurrences, not only across pieces by the same conductor, but also across different conductors.

Both attenuation and intensification may be requested through various body parts and combinations of them. For instance, "play (or sing) *piano*" can be communicated by hands, voice, posture, gaze: here are some examples:

1. a symbolic gesture: *index finger extended closing lips*, that means "be silent"
2. an iconic gesture: a *very fluid movement by left hand open palm down*
3. a codified interjection, *shhh*, a request for silence, performed (during rehearsal, obviously not in concert) to ask for "piano"
4. *neck pulled back in the shoulders*
5. *raised eyebrows*

Conversely, every modality has various signals at disposal to convey intensity indications. Within gaze signals, *raised eyebrows* are used when asking for *piano* or *delicato*, while *squeezed eyes* and *frown* ask for *forte*. Among gestures, *flat hand palm down moving like on a flat surface*, as well as *flat hand palm to musicians slightly moving towards them* convey *piano*.

Sometimes it is not a whole gesture, but simply one parameter of it (handshape, movement, location, orientation, or even the expressivity parameters of gestures, velocity, fluidity, amplitude, Hartmann et al., 2002) that conveys meanings of attenuation or intensification.

For instance,

- within direction, *movements towards players* convey “forte”, while *refraining*, *pulling back* means “piano”;
- fluidity: very *fluid movements* point at piano, *jerky movements* to forte;
- handshape: *fist* and *hand crippled as a paw* convey strength (forte), *flat* or *curve hand* convey piano.

Another interesting result concerns the semiotic devices used in the construction of intensity signals.
They may be:

1. generic codified arbitrary: Codified signals used with the same meaning as in everyday conversation: e.g., *Index finger on lips* to convey “be silent”;
2. specific codified arbitrary: Codified signals used with a specific meaning in musical performance, e.g., gestures that are used by Conductors but not by laypeople;
3. direct iconic: signals imitating some movement or some change in another modality. E.g., the Conductor’s *arms curve enlarging*, to imitate a body enflating, are an iconic gesture to convey a *crescendo*: an inflating sound. This is a case of “transmodal iconicity”, a metaphorical transposition of amplitude from a tactile and visual domain to an auditory one;
4. indirect iconic: a signal that mitates the expression of an emotion which arousal helps produce the wanted sound. E.g., *raised eyebrows* expressing anger as a prime to “forte”.
5. motoric attitude: signals imitating the movement done while producing a certain movement or the resulting sound. E.g., *squeezing eyes* to mean “sforzato”.

4. Conclusion

The use of body signals to indicate intensity in music performance is not random, rather it is systematic and shared across musicians; not only are the same signals used for the same meanings in different performances and different conductors, but the origin of those signals follows precise rules: some are iconic, and their iconicity exploits a transmodal shift from another modality to the acoustic one; some are drawn from codified signals of everyday life, others finally have become codified among musicians. Such common lexicon makes conducting a universal language allowing further intercomprehension in the universal language of music.

References

Ashley, R. (2000). The pragmatics of conducting: Analyzing and interpreting conductors' expressive gestures. In C. Woods, G. Luck, R. Brochard, F. Seddon & J. A. Sloboda (eds.) *International Conference on Music Perception and Cognition, Keele, UK,* (*2000*).

Boyes Braem, P., & Braem, T. 2004. “Expressive gestures used by classical orchestra conductors”. In Proceedings of the Symposium on The Semantics and Pragmatics of everyday Gestures, C. Mueller and R.Posner (eds), pp. 127–143. Berlin: Weidler.

Dineen, P.M. (2011). Gestural Economies in Conducting. In A.Gritten and E.King, E. (Eds) *New Perspectives on Music and Gesture*. Farnham: Ashgate.

Friberg, A. 2005. “Home Conducting: Control the Overall Musical Expression with Gestures”. *Proceedings of the 2005 International Computer Music Conference*, 479–482. San Francisco, California: International Computer Music Association.

Gritten, A. and King, E. (Eds) (2011). *New Perspectives on Music and Gesture*. Farnham: Ashgate.

Hartmann, B., Mancini, M. and Pelachaud, C. 2002. “Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis”. *Computer Animation* 2002: 111–119.

Luck, G. (2011). Computational Analysis of Conductors’ Temporal Gestures.’ In A.Gritten and E.King, E. (Eds) *New Perspectives on Music and Gesture*. Farnham: Ashgate. by Geoff Luck.

Poggi I. (2011). Music and leadership: the Choir Conductor’s multimodal communication. In M.Ichino & G.Stam (Eds.), *Integrating Gestures. The interdisciplinary nature of gestures*. Amsterdam: John Benjamins, 2011, pp.341-353.

Deriving Individual Gesture Profiles from a Multimodal Dialogue Corpus

Matthias A. Priesters, Kirsten Bergmann & Stefan Kopp
CITEC, Faculty of Technology, Bielefeld University

One of the main challenges for gesture research is that co-speech gestures are subject to substantial inter-individual differences and idiosyncrasies, regarding not only their frequency but also their type and physical form, including preferences for representation techniques or hand shapes (Bergmann & Kopp 2010) or the use of gesture space (Priesters & Mittelberg 2013). It has been suggested that verbal and spatial skills (Hostetter & Alibali 2007, Chu et al. 2014) or personality traits (Hostetter & Potthoff 2012) are factors influencing these differences (e.g., persons with higher extraversion might be likely to gesture more or to make more extensive use of space while gesturing). However, it is still unclear what constitutes individual gesturing strategies, how they are influenced by such factors, and how they distribute over speakers. Are such differences idiosyncratic, or are there larger ‘clusters’ of speakers sharing similar strategy sets? To help answer these questions, we introduce a data corpus designed to derive individual gesture profiles and to study how stable (or fluid, respectively) these are in different dialogue situations.

Building on and extending an earlier project (Lücking et al. 2013), we collected a corpus which consists of almost twelve hours of material, comprising video, audio and motion capture recordings of 62 participants. Speakers were shown a route through a town with certain sights and landmarks in a Virtual Reality (VR) environment. Afterwards they were asked to describe the route to another participant, in order to enable that person to find the same way herself later on. Importantly, each speaker was recorded in two different situations: (1) to capture speaker-specific baseline profiles, direction-givers first described a route and landmarks to a confederate who only gave minimal, controlled feedback, (2) subsequently, to capture adaptive deviations from such profiles due to dialogue context, two routers who had received differing stimuli beforehand discussed their routes or gave a joint description to a third naïve recipient. In total, 25 interactions (13 dyads, 12 triads) were recorded. Our data show speakers interacting with at least two different persons. Therefore, the data allow for investigating the influence of individual interlocutors on the communicative behavior of a speaker.

Video data were recorded from four camera angles, one for each individual speaker, one for all speakers together and one from the top. Each speaker wore a wireless microphone, recording each person’s speech on separate tracks. For motion capturing, the participants wore marker suits, in order to track the position and orientation of head, neck, shoulders, elbows and wrists. Besides basic data such as age and handedness, we also recorded participants’ personality traits according to the ‘Big Five’ inventory (McCrae & Costa 1997) and (for a part of the group) verbal and spatial cognitive skills.

Analysis and annotation of the data are currently in progress. Furthermore, the corpus is being prepared for release as a data publication, with the addition of standard-compliant CMDI metadata (Freigang et al. 2014) and a data management plan. During analysis, gestures are segmented, categorized and annotated according to their form features. Speech is transcribed and grammatically analyzed. From the motion capture data, parameters such as velocity, acceleration, position in space or distance traveled are calculated for each gesture annotation. All these data will be part of the individual gesture profiles, and will be considered in relation to speakers’ individual traits. Analyses will be continued along the following three lines:

1. Describing individual gestural behavior in gesture profiles
A speaker’s individual gesture behavior consists of a multitude of factors, such as gesture frequency, spatial extent of gestures, energy, effort or expressiveness applied to gestures, or

preferences for certain hand shapes and gesture types. These characteristics are represented in gesture profiles based on manual annotation of videos (e.g., Neff et al. 2008, Bergmann & Kopp 2010) or video recognition (e.g., Malatesta et al. 2016). Going beyond earlier approaches, we add the analysis of motion capture data to generate more precise gesture profiles. Thus, one goal of this study is to derive (probabilistic) models to describe the gesture behavior of each individual participant in the corpus data. This should not only comprise fixed preferences for parameters such as spatial locations, gesture types or frequencies, but should also be interconnected and context-dependent. That is to say, it should comprise values or preferences depending on, for example, the number of interlocutors, the dialogue context or gesture types/functions.

2. Identifying factors that underlie individual differences
While generating individual gesture profiles can provide important insights, it is also important to analyze commonalities across speakers and factors that contribute to them. One likely candidate is speakers' personality traits, others are age, gender, cultural background, or spatial and verbal cognitive skills. To address these relations, gesture profiles will be correlated with personal properties of the respective speakers to investigate whether there are correspondences between gestural parameters and personal factors. Compared across a group of speakers, this might reveal general relations between gestural and personal factors.

3. Simulating gesture profiles with animated virtual humans
Relations between gestural and personal features will be further tested by transferring them to gesture generation mechanisms for virtual humans. This allows us to test different versions of gestural behavior in human-agent interaction, for example, by preferring gestures of certain categories or specific hand shapes in certain contexts, or applying parameters derived from motion capture data to the agent's gestures: Making gestures larger, smaller, faster, more fluid, in different locations in gesture space, etc. Different settings of parameters should then be tested in dialogue with humans to ascertain whether they have an effect on how the agent is perceived, for example, if it is perceived to possess the intended personality traits.

References

Bergmann, K. & Kopp, S. (2010). Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In Kopp, S. and Wachsmuth, I. (Eds.), *Gesture in embodied communication and human-computer interaction*, pp. 182–194. Springer, Berlin/Heidelberg.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, 143(2):694–709.

Freigang, F., Priesters, M., Nishio, R., & Bergmann, K. (2014). Sharing Multimodal Data: A Novel Metadata Session Profile for Multimodal Corpora. In J. Odijk (Ed.), *Selected Papers from the CLARIN 2014 Conference*, pp. 25–35. Linköping Electronic Conference Proceedings.

Hostetter, A. B. & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95.

Hostetter, A. B. & Potthoff, A. L. (2012). Effects of personality and social situation on representational gesture production. *Gesture*, 12(1):62–83.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Malatesta, L., Asteriadis, S., Caridakis, G., Vasalou, A., & Karpouzis, K. (2016). Associating gesture expressivity with affective representations. *Engineering Applications of Artificial Intelligence*, 51:124–135.

McCrae, R. R. & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5):509–516.

Neff, M., Kipp, M., Albrecht, I., & Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph.*, 27(1):5:1–5:24.

Priesters, M. A. & Mittelberg, I. (2013). Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study. In *Proceedings of the Tilburg Gesture Research Meeting (TiGeR 2013)*.

Gesture Recognition System for Isolated Sign Language Signs

Kalin Stefanov and Jonas Beskow

KTH Royal Institute of Technology
TMH Department of Speech, Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
{kalins,beskow}@kth.se
http://www.speech.kth.se/

Abstract. This paper describes a system for recognition of isolated Swedish sign language signs for the purpose of an educational "signing game". The primary target group is children with communicative disabilities, and the goal is to offer a playful and interactive way of learning and practicing sign language signs for these children and their friends and families. Two datasets consisting of 51 signs have been recorded for a total of 7 (experienced) and 10 (inexperienced) adult signers. The signers performed all of the signs five times and were captured with a RGB-D (Kinect) sensor, via a purpose-built recording application. A recognizer based on manual features 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.

Keywords: sign language recognition, key word signing, depth sensor, real-time

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 language to support and enforce the verbal communication. As such, these communication support schemes do away with the grammatical constructs in sign language and keep only parts of the vocabulary.

While many deaf children have sign language as their first language and are able to pick it up 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. The Swedish Tivoli project, that forms the context of the system presented in this paper, aims at creating a learning environment where children can pick up 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

This paper presents a gesture recognition system that attempts to model and recognize manual features of sign language. Cooper, Holt and Bowden [1] provides comprehensive overview of the research on sign language recognition (SLR) and the main challenges. Manual features 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 by Mitra and Acharya [2].

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. The generalized topology of a hidden Markov model (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 $\sum_{j=1}^{N} a_{ij} = 1$.

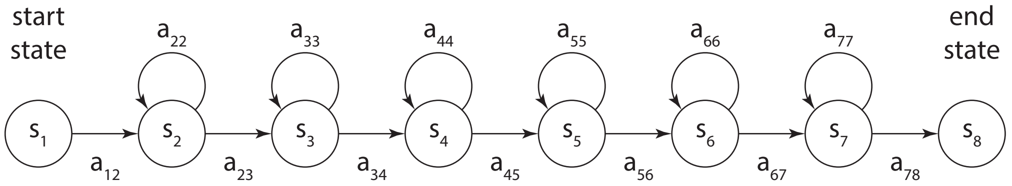


Fig. 1. 8-state left-to-right HMM used for gesture modeling.

3 Recognizer

The system is trained to perform signer independent recognition in real-time. The size of the vocabulary is 51 signs from the Swedish sign language (SSL). The vocabulary is composed of four subsets - objects, colors, animals, and attributes. The recognition rate (accuracy) in the signer dependent case is presented in Table 1. The models were trained on 4 instances of each sign for each participant and tested on the 5th instance of the sign performed by that participant. The recognition rate (accuracy) in the signer independent case is presented in Table 2. The models were trained on 30 (six different experienced signers) or 45 (nine different inexperienced signers) instances of each sign and tested on the 5 instances of the sign performed by the left out participant. All results are based on leave-one-out cross-validation procedure.

Table 1. Signer dependent results

Experienced	μ	σ	Inexperienced	μ	σ
1	92	5.1	1	84.3	4.2
2	85.5	11.1	2	92.2	6
3	84.4	7.5	3	75.7	12.1
4	**95.3**	3.3	4	94.9	1.7
5	88	14.6	5	94.5	4.2
6	87.8	7.8	6	93.3	9.4
7	80	10.8	7	89.4	5.8
			8	**95.3**	5.5
			9	94.1	6.8
			10	89.8	6.4

Table 2. Signer independent results

Experienced	1-BEST	Inexperienced	1-BEST
1	62	1	58.8
2	53.3	2	72.2
3	48	3	57.3
4	**65.5**	4	69.4
5	62.6	5	75.3
6	59.2	6	**76.1**
7	54.9	7	73.7
		8	65.5
		9	**76.1**
		10	64.7
μ	57.9	μ	68.9
σ	6.1	σ	7

As expected, the performance in the signer independent case is significantly lower than in the signer dependent case - 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 application, there is no situation where the recognizer needs to pick one sign from the full set, instead there is always the case of one out of a small number (e.g. choose one out of five objects). For this type of limited recognition tasks, accuracy will increase drastically. Furthermore, we can control the mix of signs in the game, meaning that we can make sure that signs which are confused by the recognizer never appear together, therefore, the recognition accuracy of the signer independent recognizer will not be a limiting factor in the game.

In this work each sign is modeled with the same simple HMM (see Figure 1). More complex and different models for each sign will be considered in future work. Further improvements are expected by introducing adaptation of the HMMs based on a small set of signs from the target signer that could be collected during an enrollment/training phase in the game.

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 recording the experienced signers dataset.

References

1. H. Cooper, B. Holt, R. Bowden, Sign language recognition, in: T. B. Moeslund, A. Hilton, V. Kroger, L. Sigal (Eds.), Visual Analysis of Humans, Springer, 2011, pp. 539-562.
2. S. Mitra, T. Acharya, Gesture recognition: A survey, IEEE Transactions on Systems, Man, and Cybernetics 37 (3) (2007) 311-324.

Time alignment between laughter and laughable

Ye Tian[1], **Chiara Mazzocconi**[2] **Jonathan Ginzburg**[2,3]
[1]Laboratoire Linguistique Formelle (UMR 7110)
& [2]CLILLAC-ARP (EA 3967) & [3]Laboratoire d'Excellence (LabEx)—EFL
Université Paris-Diderot, Paris, France
tiany.03@googlemail.com

Several authors have advanced claims about what laughter is about based exclusively on its temporal sequence, therefore resolving the referent of the laughter as the utterance or the event immediately preceding or following it (Provine, 1993; Vettin and Todt, 2004). Analysis of the laughter's preceding elements has subsequently lead to the conclusion that laughter is rarely related to "funny" stimuli because most frequently it follows "banal" comments (Provine, 1993). Another frequently cited piece of data is that laughter *punctuates speech* (Provine, 1993; Provine and Emmorey, 2006) i.e., it never interrupts phrases, neither by the speaker nor by the listener. The aim of our current study is to verify the validity of these popular assumptions and claims, by analysing data from a French and Chinese dialogue corpus.

We argue that previous studies, not attempting at integrating laughter in an explicit semantic/pragmatic module, have overlooked a detailed analysis of the laughable. Following the account of (Ginzburg et al., 2015) we argue that laughter is a gestural event anaphor, whose meaning contains two dimensions: one dimension about the arousal and the other about the trigger or the laughable. In line with (Morreall, 1983) we think that laughter signals a "positive psychological shift", and the "arousal" dimension signals the amplitude in the shift[1] The positive psychological shift is triggered by an appraisal of an event - the laughable l, and the second dimension communicates the type of the appraisal. Ginzburg et al. (2015) propose two basic types of meaning in the laughable dimension: the person laughing may express her perception of the laughable l as being incongruous, or just that l is enjoyable (playful)[2]. We propose that in addition, certain uses of laughter in dialogue may suggest the need for a third possible type: expressing that l is a socially close ingroup situation[3].

In this paper we address these questions:

1. Does Laughter always follow its laughable? If not, does the laughter-laughable alignment differ among different types of laughters?

2. Does laughter interrupt speech?

1 Materials and Method

We analyzed a portion of the DUEL corpus (Hough et al., 2016): two couples interacting in a natural, face-to-face, loosely task-directed dialogue, both in French and Mandarin Chinese (2 pairs x 2 languages), having a total of 657 laughter events analysed in relation to their laughable over a total of 160mins. **Audio-video coding of laughter**: each video was observed until a laugh occurred sharing the same criteria utilised in (Nwokah et al., 1994; Apte, 1985; Ekman and Friesen, 1975). The coder detected the onset and offset of laughter, and conducted a multi-layer analysis as illustrated in Figure 1. **Audio-video coding of laughable**: we consider as the laughable the event which, after appraisal, produces a positive psychological shift in the laugher. We distinguish three different kinds of laughable types: described events, metalinguistic stimuli and exophoric events (see Fig.1 for definitions).

2 Results

	French	Chinese
Dialogue.dur	77min	85min
mean utterance.dur	1.8sec	1.5sec
No. laughter	436	221
laughter.dur	1.9s (sd .97)	1.4s (se .53)
No. laughable	256	158
laughable.dur	2.7s (sd 1.5)	2.8s (sd 2.1)
No.laughter per laughable	1.7	1.4

Table 1: Data summary

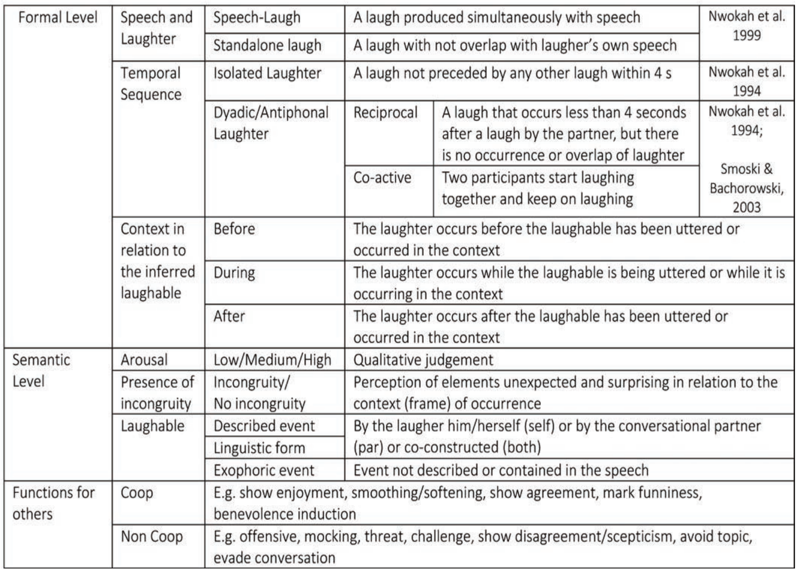
2.1 Does laughter always follow the laughable?

In both Chinese and French, on average, laughter starts *during* rather than after the laughable, and

[1] The amplitudes in the shift depend on both the trigger itself and on the individual current information/emotional state. It is important to point out that laughter does not signal that the speaker's current emotional state is positive, merely that there was a shift which was positive. The speaker could have a very negative baseline emotional state (being very sad or angry) but the recognition of the incongruity in the laughable or its enjoyment can provoke a positive shift (which could be very minor).

[2] Ironic displays of laughter convey exactly the opposite meaning:i.e., no positive shift has been triggered

[3] We defer to forthcoming papers a more detailed discussion on theories of laughter



Formal Level	Speech and Laughter	Speech-Laugh	A laugh produced simultaneously with speech		Nwokah et al. 1999
		Standalone laugh	A laugh with not overlap with laugher's own speech		
	Temporal Sequence	Isolated Laughter	A laugh not preceded by any other laugh within 4 s		Nwokah et al. 1994
		Dyadic/Antiphonal Laughter	Reciprocal	A laugh that occurs less than 4 seconds after a laugh by the partner, but there is no occurrence or overlap of laughter	Nwokah et al. 1994; Smoski & Bachorowski, 2003
			Co-active	Two participants start laughing together and keep on laughing	
	Context in relation to the inferred laughable	Before	The laughter occurs before the laughable has been uttered or occurred in the context		
		During	The laughter occurs while the laughable is being uttered or while it is occurring in the context		
		After	The laughter occurs after the laughable has been uttered or occurred in the context		
Semantic Level	Arousal	Low/Medium/High	Qualitative judgement		
	Presence of incongruity	Incongruity/ No incongruity	Perception of elements unexpected and surprising in relation to the context (frame) of occurrence		
	Laughable	Described event	By the laugher him/herself (self) or by the conversational partner (par) or co-constructed (both)		
		Linguistic form			
		Exophoric event	Event not described or contained in the speech		
Functions for others	Coop	E.g. show enjoyment, smoothing/softening, show agreement, mark funniness, benevolence induction			
	Non Coop	E.g. offensive, mocking, threat, challenge, show disagreement/scepticism, avoid topic, evade conversation			

Figure 1: Laughter coding parameters

finishes after the laughable. The distribution varies over a wide range and laughter is shown to occur both during as well as after or before with a gap up to 10 sec (see Fig.2). Figure 3 shows some of the possible patterns we observed.

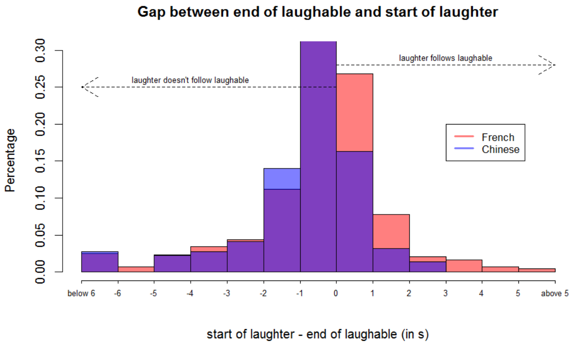


Figure 2: Gap between laughable and laughter

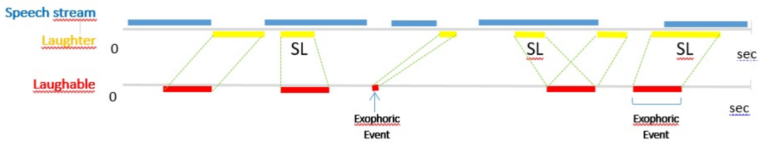


Figure 3: Temporal misalignment speech stream, laughter and laughable

2.2 Does laughter-laughable alignment differ among different "types" of laughables and laughters?

On average, there are more self-produced than partner-produced laughables, supporting the idea that speakers laugh more often than the addressees. Laughters about a partner-produced laughable start later than those about a self-produced laughable. Laughter frequently overlaps with speech (36% in French, 47% in Chinese) Speech laughters overlap with the laughable more than laughter bouts. 52% of speech laughters in French and 70% in Chinese overlap with the laughables. In comparison, 33% of laughter bouts in French and 34% in Chinese overlap with the laughable. Notice that not all speech laughters overlap with the laughable, suggesting that often, laughter that co-occurs with speech is not about the co-occurring speech (47.8% in French and 30% in Chinese).

2.3 Does Laughter interrupt speech?

Yes. 51.8% of laughter bouts in French and 56.7% of laughter bouts in Chinese start during the partner's utterances (not necessarily laughables). Laughter can also occur in utterance-medial position (5% in French and 8.6% in Chinese, significantly higher than zero: French $\chi^2(1)=12.3$, $p=.0004$; Chinese $\chi^2(1)=10.5$, $p=.001$). Most of these interruptions at not at phrase boundaries.

3 Discussion and conclusions

Our results show that laughter displays a rather free alignment in respect to its laughable and, contrary to popular belief, only 30% of laughter occur immediately after their laughable. Laughter can occur during, long before or long after (up to 10 s) its laughable. Therefore invalidating any attempt to infer what the laughter is about exclusively based on adjacency. Interestingly, speech-laughter often does not overlap with its laughable, suggesting that frequently laughs are not about the co-occurring speech. Laughter-laughable alignment may differ depending on the different "types" of laughable and laughter. Specifically, laughters about a partner-produced laughable (audience laughter) start later than those about a self-produced laughable (speaker laughter).

Our data also invalidate the claim that laughter punctuates speech (Provine, 1993), indeed laughter does interrupt speech. We often laugh when others are speaking (half of all laughter bouts) and occasionally we insert stand-alone laughters mid-sentence (less than 10%). Moreover, very frequently laughter overlaps speech (around 40% of all laughters).

Cross-linguistically the patterns are similar, except that in Chinese, laughters are more likely to overlap with the laughable than in French. This suggests that while certain aspects of laughter behaviour is influenced by culture/language, generally we use laughter similarly in interaction.

References

Mahadev L Apte. 1985. *Humor and laughter: An anthropological approach*. Cornell Univ Pr.

Paul Ekman and Wallace V Friesen. 1975. Unmasking the face: A guide to recognizing emotions from facial cues.

Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*, University of Amsterdam.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *10th edition of the Language Resources and Evaluation Conference*.

John Morreall. 1983. *Taking laughter seriously*. SUNY Press.

Evangeline E Nwokah, Hui-Chin Hsu, Olga Dobrowolska, and Alan Fogel. 1994. The development of laughter in mother-infant communication: Timing parameters and temporal sequences. *Infant Behavior and Development*, 17(1):23–35.

Robert R Provine and Karen Emmorey. 2006. Laughter among deaf signers. *Journal of Deaf Studies and Deaf Education*, 11(4):403–409.

R. R. Provine. 1993. Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology*, 95(4):291–298.

Julia Vettin and Dietmar Todt. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.

A repertoire of multimodal signals communicating certainty and uncertainty

Laura Vincze and Isabella Poggi
Department of Philosophy, Communication and Visual Arts
University Roma Tre – Rome, Italy

1. A repertoire of body signals of certainty and uncertainty

When having to communicate a belief towards whose truth-value we are not highly committed, our body generally 'tells' that we are being uncertain. Same for the beliefs we are certain of: our body multimodally communicates our certainty. Research investigating the degree of confidence the speaker has in his statement has condensed in the field of "epistemicity". Epistemicity can be defined as "the degree of confidence the speaker has in his or her statement". In other words, it refers to the degree of commitment that a speaker has in the truth-value of a proposition.

Traditional studies on epistemicity (Chafe 1986; Dendale and Tasmowski 2001; De Haan 2001; Kärkkäinen 2003; Biber & Finegan 1989; Biber 2004; Hoye 2008; Cornillie 2009; Heritage 2012a, b; Heritage 2013; Marín Arrese 2014; Bongelli & Zuczkowski 2008, among others) have focused predominantly on how languages use morphosyntactic and discoursal features to communicate epistemic stance, that is, the epistemic attitude of the Speaker towards the information s/he is delivering. Not so much attention has been dedicated to the range of multimodal strategies speakers have at their disposal to communicate the uncertainty or certainty of their beliefs. In the epistemicity literature, studies focusing on the role of facial expression, prosody and gesture in the communication of speaker's epistemic stance (Borràs-Comes et al. 2011; Roseano et al. 2014; Dijkstra et al. 2006; Bitti et al. 2014; Mondada 2013) are an exception rather than the rule. Moreover, the few studies devoted to the analysis of the speaker's *overall (un)certain stance* focus on "traditional" signals such as *head nod* as a certainty signal (Roseano et al. 2014), and *head tilt* (Heylen 2005) or facial expression (*raised eyebrows* and *lowered mouth angles*) to signal uncertainty (Ricci Bitti et al. 2014). The present study aims to widen the traditional perspective adopted by studies in the fields of epistemicity by investigating speakers' *epistemic stance* not so much as it is expressed in words or syntax, but through bodily signals. The goal of our work is to outline a repertoire of signals of certainty and uncertainty performed through body behaviours, namely by gestures, gaze, facial expression, head movements, and postures.

2. Corpus and Method

In order to investigate body signals of (un)certainty, we collected a corpus of 100 video-abstracts of British Medical Journal[1] in which medical researchers illustrate orally their findings published in the journal, with the aim of a more rapid dissemination to peers. The speakers are quite gender-balanced, amongst a total of 62 speakers, there are 28 females and 34 males. All of them, with the exception of 3 males and 3 females, are native English Speakers. Signals in various modalities informing on the speaker's certainty or uncertainty towards the communicated beliefs were singled out and analysed: gestures, head movements, body movements, facial expressions, gaze items. Each signal was described as to its parameters of shape and movement, their verbal and body context was annotated, thus 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 common meaning (for example, certainty of a statement can be reinforced by a head nod, an eye closure, or a headshake).

[1] The British Medical Journal invites authors of research articles to explain their work in a short video abstract (4-5 min in average).

3. Results

This work resulted in a repertoire of signals devoted to convey certainty and uncertainty in several body modalities. Within it, both cases of synonymy and of polysemy were found. Two cases of synonymy, that is, two different signals both used to convey the speaker's certainty are for example the *shoulder shrug*, often accompanied by *a palm up open hand gesture* (Müller, 2004; Müller & Cienki, 2008), and the *intensity head shake*, often accompanied by an *intensity eye closure*.

But also the reverse of synonymy can be the case, polysemy: the same signal conveying quite different meanings, some of which are connected to certainty or uncertainty, while others are not. This occurs with the *shoulder shrug*, that may convey several meanings: from *obviousness* of a state of affairs (Debras & Cienki 2012) to *lack of knowledge* (Jokinen & Allwood 2010), to *unimportance* or finally *helplessness*. When finding such cases of polysemy, we tried to find a core of meaning that is common to the apparently divergent meanings of that same signal, whether or not connected with (un)certainty.

For example, the *shoulder shrug* may assume, in different contexts, quite diverse meanings, five of which already found by Debras & Cienki (2012), and one in our corpus: uncertainty, ignorance, obviousness, indifference, unimportance, impotence. Out of them, only the first three are linked to epistemic stances, while the last three are more pertinent to goals than to beliefs. Moreover, uncertainty and ignorance seem opposite to obviousness; the first two having to do with the uncertain or unknown (Bongelli & Zuczkowski 2008), and the third being more a signal of certainty. Notwithstanding this, all six meanings of the *shrug* share a common semantic component of "discharge of responsibility", non-intervention. As for the meanings of "unimportance" and "indifference", I do not intervene because I do not have the goal to change things, they are not relevant for me; for "impotence" I do not because I have no power over it, so it would be wasted time trying to do something. "Ignorance" is a form of epistemic impotence, and I do not intervene because I really do not know; in "uncertainty" I do not intervene because, not being certain of something, I cannot take on responsibility about its truth value; in "obviousness" there is no point to intervene, since it is certain and clear to everybody.

Another recurrent body signal conveying certainty is the *headshake*. Headshakes in our corpus convey the following three meanings: "intensification" (Kendon 2002), "absolute inclusivity" (McClave 2000; Heylen 2005) and "negation", all possibly associated to certainty.

Intensification headshakes usually co-occur with words such as *extremely, very,* or with other intensification multimodal signals such as *eye-closure* (McClave 2000; Vincze & Poggi 2011), that all convey a high degree of certainty.

By means of *headshakes*, the speaker can mark his high degree of certainty both in positive ("I am certain this is so") and in negative ("I am certain this is not so"). While the former type of headshake co-occurs with lexical affiliates such as *very, particularly, extremely*, the latter co-occurs with verbal negation.

Besides conveying an intensified belief (with a high degree of certainty), *headshakes* also convey the meaning of inclusivity; that is, they mean that the properties or events being stated are considered as extended to many or all possible cases. In this meaning, headshakes co-occur with pronouns such as *any-; every-*.

These forms of intensification and absolute inclusivity communicate the certainty of the speaker, possibly aiming at persuasive goals through increasing certainty in the listener as well. Actually, our hypothesis is that stating a high level of something may contribute, in many cases, to state a high certainty of what is stated (Poggi & D'Errico, 2016).

Just like when communicating certain beliefs we multimodally signal our certainty, in the same vein, when we communicate a belief whose truth-value we are not highly committed to, our body generally attenuates the strength of our assertion, meta-communicating our uncertainty.

In our corpus we singled out the following uncertainty signals: *approximation hand gestures, head* *tilts, facial expressions conveying speaker's dissatisfaction concerning his own assertion.*

Let's take the first mentioned signals, *approximation gestures*. As defined in previous works (Channel, 1992; Poggi & Vincze, 2012; Vincze et al., 2012), a speaker is approximate either because he does not have the power to be precise (the speaker lacks precise knowledge about quantities, therefore he is uncertain about which, in a range of different possible quantities, is the true one) or because the speaker does not have the goal to be precise (the speaker knows the precise quantities but, for the purposes of the ongoing interaction, he considers it unnecessary to be precise). In our study we distinguish between the two causes of approximation and consider the former only.

Head tilts are traditionally known to convey speaker's uncertainty. In our corpus they co-occur with lexical markers of uncertainty such as *unlikely* or *perhaps*. They may come alone or accompanied by a dissatisfied facial expression, for instance with *protruded lips* and/or *lowered lip corners*, signalling at a first level that the speaker does not like very much what he is saying, and at a second level that he does not very much believe the stated hypothesis.

4. Conclusion

The presented study aims at an overview of body signals of certainty and uncertainty in Anglo-American cultures and goes in the direction of creating a repertoire of them; the signals found in the analyzed corpus will later be investigated in other corpora of different activity types and different cultures.

References:

Aikhenvald, Alexandra (2004). *Evidentiality*. Oxford University Press: Oxford.

Bongelli, R. Zuczkowski, A. (2008). *Indicatori linguistici percettivi e cognitivi*. Aracne, Roma

Debras, Camille, & Alan Cienki. 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 Confernece on Social Computing (SocialCom)*. IEEE, 2012.

Heritage, J. (2012a). Epistemics in Action: Action Formation and Territories of Knowledge. *Research on language and social interaction, 45(1)* , 1-29.

Heritage, J. (2012b). The Epistemic Engine: Sequence Organization and Territories of Knowledge. *Research on Language and Social Interaction, 45 (1)* , 30-52.

Heritage, John (2013). Action formation and its epistemic (and other) backgrounds. In: Discourse Studies 15(5) 551-578, Sage

Heylan, D. "Challenges ahead. head movements and other social acts in conversation." AISB, 2005.

Jokinen, K., & Jens A. "Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging." *Culture and computing*. Springer Berlin Heidelberg, 2010. 55-70.

Kärkkäinen, Elise, 2003. *Epistemic Stance in English Conversations. A Description of its Interactional Functions, with a Focus on I Think*. Amsterdam, Benjamins

Kendon, Adam. "Some uses of the head shake." *Gesture* 2.2 (2002): 147-182.

McClave, Evelyn Z. "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, Sage, (2013)

Poggi, Isabella, and Laura Vincze. "Communicating vagueness by hands and face." *Proceedings of the ICMI Workshop on Multimodal Corpora for Machine Learning, Alicante*. 2012.

Poggi I, & D'Errico F. (2016) Finding Mussolini's charisma in his multimodal discourse. In F.Paglieri, L.Bonelli & S.Felletti (Eds.) *The Psychology of Argument. Cognitive Approaches to Argumentation and Persuasion*. London: College Publications.

Roseano, Paolo; González Montserrat, Borràs-Comes, Joan & Prieto, Pilar (2014) Communicating Epistemic Stance: How Speech and Gesture Patterns Reflect Epistemicity and Evidentiality. In *Discourse Processes*. DOI:10.1080/0163853X.2014.969137

Cornillie, Bert (2010). An interactional approach to epistemic and evidential adverbs. In Diewald, Gabriele & Smirnova, Elena (eds.), *Linguistic Realization of Evidentiality in European Languages*. Walter de Gruyter: Berlin / New York, 309,330.

Dijkstra, Christel; Krahmer, Emiel & Swerts, Marc (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*.

Marín Arrese, Juana (2011). Epistemic legitimizing strategies, commitment and accountability in discourse. *Discourse Studies* 13(6), 789-797.

Vincze, Laura, Isabella Poggi. "Communicative functions of eye closing behaviours." *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*. Springer Berlin Heidelberg, 2011. 393-405.

Vincze, Laura, Isabella Poggi, and Francesca D'Errico. "Vagueness and dreams: analysis of body signals in vague dream telling." *Human Behavior Understanding*. Springer Berlin Heidelberg, 2012. 77-89.

Can co-speech gesture change the perception of ambiguous motion events?

Experimental evidence from Italian

Bjørn Wessel-Tolvig & Patrizia Paggio

University of Copenhagen

How sensitive are we to information provided by co-speech gestures when interpreting ambiguous motion events? Recent findings extent the tight integration of speech and gesture, as observed in production [1, 2], to also include comprehension [3, 4]. This integration implies that listeners integrate information in gestures in the understanding of the utterance.

We present an experimental judgment task in which we investigate the effect of gestural information on listeners' interpretation of different types of ambiguous motion constructions in Italian.

One of the major limitations for verb-framed languages is the boundary-crossing constraint [5, 6], which predicts that movement across a spatial boundary cannot be lexicalized with a manner verb and a preposition, e.g. as done in the English '*roll into*'. Boundary-crossing motion must be expressed in a different way, e.g. a path verb like '*enter*' since the prepositional system is inherently locative and these prepositions therefore do not encode the directionality needed to express translational motion across spatial boundaries [7].

Romance languages, and among them Italian, are generally classified as verb-framed languages [8, 9]. However, recent theoretical debate suggests that speakers of Italian may disregard the constraint and express boundary-crossing movement with certain types of manner verbs and complex PP combinations [10, 11]. An expression like '*Il pallone rotola fuori dalla stanza*' can thus be read as '*the ball rolls out of/outside the room*'. The reading often depends on contextual inference or pragmatic clues [12].

Before testing the effect of co-speech gestures on the interpretation of ambiguous motion event expressions, a questionnaire (109 Italian participants) was conducted to verify the interpretation of manner verb + PP constructions involving different types of manner verbs (directional manner verbs and pure manner verbs). The results confirm the existence of boundary-crossing interpretation for certain types of Italian manner verb + PP constructions.

Based on these results, we created video materials in which an Italian native speaker expressed the same motion events as in the first task, but this time with different co-speech gestures. Half of the verb + PP constructions in each group were accompanied by directional (path) gestures, and the other with locative (manner) gestures. Italian participants (n = 103) were asked to judge the utterances whether they denoted boundary or non-boundary-crossing movement.

The results show that co-speech gesture can change the perception of the events. A locative manner gesture can change the default interpretation of verb + PP constructions that allow boundary-crossing readings to be interpreted as locative, and directional path gestures can change the interpretation of verb + PP constructions that do not allow boundary-crossing interpretation to be perceived as boundary-crossing.

To summarize, 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 perception of events and thus ‘override’ default meaning expressed only in speech.

[1] D. McNeill, *Gesture & Thought*. Chicago: The University of Chicago Press, 2005.

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

[3] S. Kelly, A. Özyürek, and E. Maris, "Two Sides of the Same Coin: Speech and Gesture Mutually Interact to Enhance Comprehension," *Psychological Science,* vol. 21, pp. 260-267, February 1, 2010 2010.

[4] H. Holle and T. C. Gunter, "The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence," *Journal of Cognitive Neuroscience,* vol. 19, pp. 1175-1192, 2007/07/01 2007.

[5] R. A. Berman and D. I. Slobin, *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Psychology Press, 1994.

[6] J. Aske, "Path predicates in English and Spanish: A closer look," in *Berkeley Linguistic Society 15*, Berkeley, USA, 1989.

[7] Ş. Özçalışkan, "Ways of crossing a spatial boundary in typologically distinct languages," *Applied Psycholinguistics,* vol. FirstView, pp. 1-24, 2013.

[8] A. Pavlenko, *Thinking and speaking in two languages*. Clevedon, UK: Multilingual Matters, 2011.

[9] Z. Han and T. Cadierno, *Linguistic relativity in SLA: Thinking for speaking*. Bristol: Multilingual Matters, 2010.

[10] F.-E. Cardini, "Grammatical constraints and verb-framed languages: The case of Italian," *Language and Cognition,* vol. 4, pp. 167-201, 2012.

[11] R. Folli, "Complex PPs in Italian," in *Syntax and Semantics of Spatial P.*, A. Asbury, J. Dotlacil, B. Gehrke, and R. Nouwen, Eds., ed: John Benjamins Publishing Company, 2008, pp. 197-221.

[12] J. Beavers, B. Levin, and S. Wei Tham, "The typology of motion expressions revisited," *Journal of Linguistics,* vol. 46, pp. 331-377, 2010.