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

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

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

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

Abstract

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

1 Introduction

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

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

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

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

2 Background

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

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

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

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

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

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

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

3 Velocity, acceleration and jerk

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

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

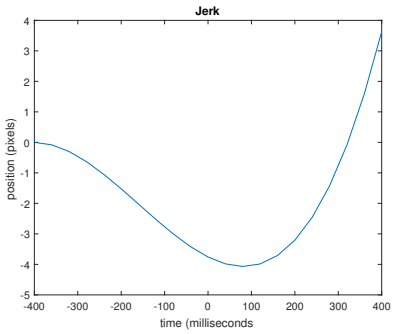

Figure 1: Jerk in an idealised nod. The figure depicts the relative position of the head in a time period of 800 milliseconds. Initially, at -400 ms, the head is at rest in position 0. Due to a negative (downward) acceleration it moves a few pixels down, but the positive, constant (43.74 $pixels/s^3$)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

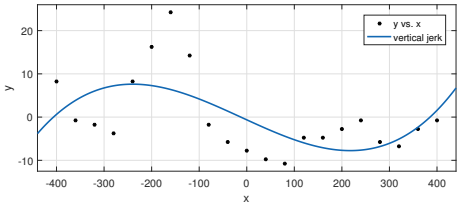

Figure 2: Jerk during a nod computed from data produced by the Anvil Facetracker. The figure depicts the relative vertical position of the head in a time period of about 800 milliseconds, or 21 frames (assuming a frame rate of 25 frames/s). Points on the curve approximate the relative position of the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

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

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

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

4 Data, training and test setup

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

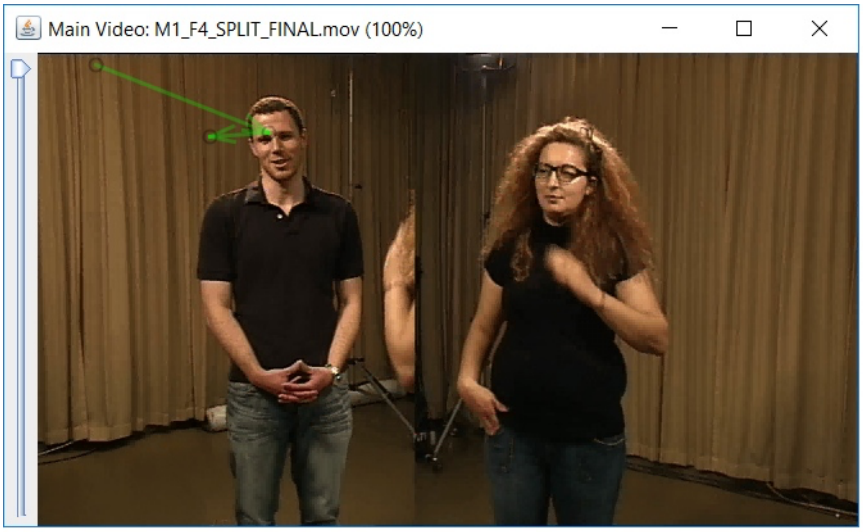

Figure 3: A frame in the midst of an HeadOther movement. The arrow indicates the strength and direction of the jerk movement feature

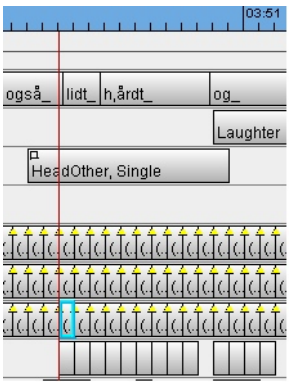

Figure 4: Area in Annotation window showing from top to bottom the manual annotation for the HeadOther movement in the previous figure, three tracks of frame-by-frame annotations for velocity, acceleration and jerk, and the annotation predicted by the SVM model. It can be seen that the Laughter annotation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

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

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

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

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

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

other was set aside as test data.

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

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

5 Results and discussion

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

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

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

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

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

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

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

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

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

6 Conclusions

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

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

Table 3: Number of head movements identified by the system

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

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

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

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

References

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

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Jean-Claude Martin, Patrizia Paggio, Peter Kuehnlein, Rainer Stiefelhagen, and Fabio Pianesi, editors, *Multimodal Corpora for Modelling Human Multimodal Behaviour*, volume 41 of *Special issue of the International Journal of Language Resources and Evaluation*, pages 273–287. Springer.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén, Kristiina Jokinen, and Costanza Navarretta. 2010. The NOMCO multimodal nordic resource - goals and characteristics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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

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

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

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