Comparing cascaded LSTM architectures for generating head motion from speech in task-oriented dialogs

Duc Canh Nguyen, Gérard Bailly, Frédéric Elisei

Comparing cascaded LSTM architectures for generating head motion from speech in task-oriented dialogs

Duc-Canh Nguyen, Gérard Bailly & Frédéric Elisei

GIPSA-Lab, Grenoble-Alpes Univ. & CNRS, Grenoble, France,
firstname.lastname@gipsa-lab.fr

Abstract. To generate action events for a humanoid robot for human robot interaction (HRI), multimodal interactive behavioral models are typically used given observed actions of the human partner(s). In previous research, we built an interactive model to generate discrete events for gaze and arm gestures, which can be used to drive our iCub humanoid robot [19, 20]. In this paper, we investigate how to generate continuous head motion in the context of a collaborative scenario where head motion contributes to verbal as well as nonverbal functions. We show that in this scenario, the fundamental frequency of speech (F0 feature) is not enough to drive head motion, while the gaze significantly contributes to the head motion generation. We propose a cascaded Long-Short Term Memory (LSTM) model that first estimates the gaze from speech content and hand gestures performed by the partner. This estimation is further used as input for the generation of the head motion. The results show that the proposed method outperforms a single-task model with the same inputs.

Keywords: Head motion generation; Human interactions; Multi-tasks learning; LSTM; Human-robot interaction

1 Introduction

Human interactions are steered by complex and multimodal sensorimotor loops [25]. In order to provide artificial agents such as virtual avatars or social robots with the ability to communicate and cooperate smoothly with humans, we need a multimodal behavioral model which can capture the co-variations of joint verbal, co-verbal and non-verbal cues of social interactions. These cues include speech, gaze, hand gestures, body postures, etc. Coordination between these cues within and between conversational partners is of major importance for shaping communicative functions as well as monitoring interpersonal and group relations. Head motion contributes to multiple functions such as visual attention, emotional display, back-channeling and is influenced by multiple social, physiological and cognitive factors.

We analyze here head motion data of a human subject involved in a face-to-face cooperative interactive game (see section 3) that requires verbal communication and visual attention. We challenge the problem of generating continuous head movements from speech activities and gestures of both partners. We will show how the exploitation of the main causal relations between speech, gestures, gaze and head motion into the modeling architecture benefits to both prediction accuracy and coordinative structures.

2 State of the art

2.1 Head motion, gaze and speech

The study of human eye-head coordination during orienting movements to targets has a long history [10]. This coordination is influenced by numerous factors including the nature of the target, its position in the field of vision and with respect to the previous fixation, etc. Head motion also contribute to active listening: it complements binaural cues [4] and has been shown to enhance automatic source diarization and localization [16]. Head motion is also important to acknowledge or replace verbal back-channels (e.g., nodding for acknowledging or shaking for signaling doubt), but also for many aspects of human communication. Munhall et al. [18] showed that vision of head motion improves speech perception. Graf et al. [9] demonstrated that the timings of head motion and the prosodic structure of the text are consistent and suggest that head motion is useful to segment the spoken content. Yehia et al [24] notably evidenced high correlation between head motion, eyebrow movements and the fundamental frequency (F0) of speech. Head motion also provides useful information about the mood of the speaker [5].

2.2 Predicting head motion

Rule-based systems are common methods to monitor human interactions. For example, Liu et al [14] proposed to generate head pan by analyzing utterance structure and identifying backchannels, while head tilt was depending on phrase length. Cassell et al [6] build a conversational system which coordinated facial expression, eye gaze, head and arm motion. A similar Nonverbal Behavior Generator system was proposed by Lee and Marsella [12] that associates multimodal patterns with given communication functions. Thorisson [22] used a finite state machine to describe events of interaction scenario with pre-conditions and post-actions in different hierarchical layers. However, hand-crafted rules are difficult to handle once considering the many factors conditioning the multimodal behaviors such as emotion, task, personality, etc.

Machine learning techniques have been proposed to map functions with behaviors. For example, Busso et al [5] proposed to use Hidden Markov Models (HMM) to drive head motion from prosodic features. Ben Youssef et al [2] used articulatory features to drive head motion synthesis. Another HMM-based framework to generate body movement from prosody was proposed by Levin [13]. Ding

et al [8] also trained an HMM to generate head and eyebrow movements. Mariooryad et al [15] further explore dynamic Bayesian networks (DBN) for coupling speech with head and eyebrow movements. More recently Sadoughi et al [21] introduced latent variables to consider speaker intentions.

Recently, Recurrent Neural Networks (RNN) have been shown to outperform statistical models in sequence recognition and generation. Gated recurrent units (GRUs) and Long-Short Term Memory (LSTM) cells have been introduced to cope with long-term temporal dependencies. These cells basically add gates to inputs and outputs (and thus the ability to keep activations over short as well as long periods) of the basic processing units that perform the non-linear mapping.

Few works have been using LSTM to model human machine interaction. For example, Alahi et al [1] used LSTM with social pooling of hidden states which combines the information from all neighboring states to predict human trajectories in crowded space. Haag et al [11] proposed Bidirectional LSTM with stacked Bottleneck feature to improve the quality of head motion generation.

In this paper, we present a multimodal behavioral model to generate the head motion of an instructor during a collaborative task with a manipulator. We proposed a cascaded multitask learning method, where gaze prediction is considered as an intermediary task for further improving head motion generation. The results can be used partially to drive multimodal interactive behaviors of a humanoid robot.

3 Interactive Data

The dataset used as interactive data in this paper has been collected by Mihoub et al [17]. This face-to-face interaction involves an instructor and a manipulator who performed a collaborative task called "put that there". The experimental setting is shown in Figure 1. In this scenario, the manipulator will move cubes following guidance of the instructor as he does not know the source and target positions of the cubes . Conversely, the instructor knows the locations – delivered by a computer program via a tablet that is only visible from the instructor – but is not able to move the cubes. The task requires that they share knowledge and coordinate their sensorimotor abilities.

The data analyzed here was collected with one instructor interacting with 3 successive manipulators. Each manipulator plays 10 games. Each game consists in moving 10 cubes from a cube reservoir close to the manipulator – where 16 cubes are randomly arranged in 2 rows – to a target 8x8 chessboard located between the interlocutors. The interactive data sums up to 30 minutes, with mean duration of a game close to 80 seconds.

The interactive data were monitored by motion capture – including head motions, gestures and eye tracking – in synchrony with speech and were resampled at 25 Hz. Motion and speech data were annotated semi-automatically with Elan [23] and Praat [3]. The total data observations include finally 3 continuous motion of the instructor's head (converted to Euler angles: pitch (H1), roll (H2), and yaw (H3)) and 5 discrete variables:

– IU: Interactive Units correspond to sub-tasks and pace joint activities. We distinguish between 6 different activities: get information from tablet, find the cube to be moved, point to the cube, indicate target position, check the manipulation and validate the target position of the cube.
– SP: the instructor's speech is segmented according to 5 speech values: manipulated cube, reference cube, relative positioning, else and none
– MP: the manipulator's arm gestures are segmented into the following 5 strokes: go-to-rest, grasp, move, put the cube and else.
– GT: the instructor's arm gestures are segmented into the following 5 strokes: rest, point the manipulated cube, the reference cube vs. the target position and else.
– FX: we distinguish between 5 regions of interest of the instructor's gaze: manipulator's face, reservoir, task space (c̄hessboard), reference cube and else.

Fig. 1. First-person view of the interaction captured from the instructor's head-mounted scene camera. At a game onset, the cube reservoir close to the manipulator is full. The instructor then ask the manipulator to put certain cubes at certain places of the chessboard: at the center at onset then left/right/on top/at the bottom of cubes already released. The circle features the point of interest of the current eye fixation.

4 Multimodal interactive behavioral models for continuous variables

In previous research, Mihoub [17] and Nguyen et al [20] built multimodal behavioral models which are able to generate GT and FX given input streams SP and MP. At training stage, all of discrete streams (IU, GT, FX, SP and MP) are available, while in generating stage, only SP and MP are observed.

In this work, we investigate interactive models to generate continuous variables - head motions of the instructor head (H1, H2, H3) with the same observed input SP and MP.

4.1 Analyzing data

Canonical Correlation Analysis (CCA) is often used to measure the interdependence between two sets of sequential data with different or equal feature dimensions [7]. The basic idea of CCA is to find optimal linear combinations of features (so-called canonical variables) of each stream that maximize the correlation between canonical variables.

Using CCA, we computed the correlations between each modality (IU, GT, FX, SP, MP and F0) and head motions (H1, H2, H3). Figure 2 displays the mean correlations of this analysis. For all of angles, the highest correlations are with IU and FX. The pitch angle (H1) exhibits the highest mean correlation (0.7) while the others are about 0.5. This can be explained by the fact that FX on face, source's cube space, tablet are well separated in pitch direction while the roll and yaw (H2, H3) are actually not separated into different azimuthal regions. The least correlated feature with head motions is F0. Therefore, for this specific collaborative task, F0 is not sufficient to accurately predict head motion. This is expected since speech chunks partly refers to movable regions of interest in the visual scene that are intrinsically referred via non verbal signals such as gaze.

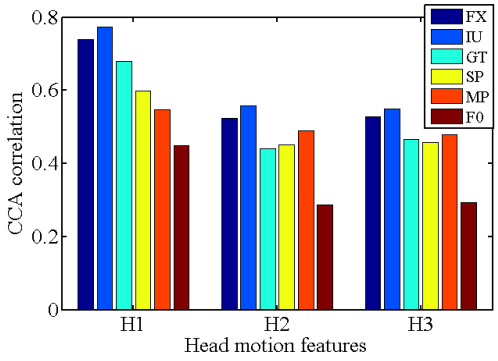


Fig. 2. Correlation of CCAs between each of H1, H2, H3 and FX, IU, GT, SP, MP and F0

We compare here the performance of mainly three different models:

Baseline. The baseline model for generating head motion uses one LSTM layer with linear activations to generate directly H1,H2,H3 as shown in Figure 3(a). This model uses the same inputs than the DBN proposed by Mihoub [17], i.e. the observed variables (SP and MP).

Control. Based on CCA analysis results (IU and FX have higher correlation with H1), the control model uses FX as an additional input feature as illustrated in Figure 3(b). Our generation models will compete with this control

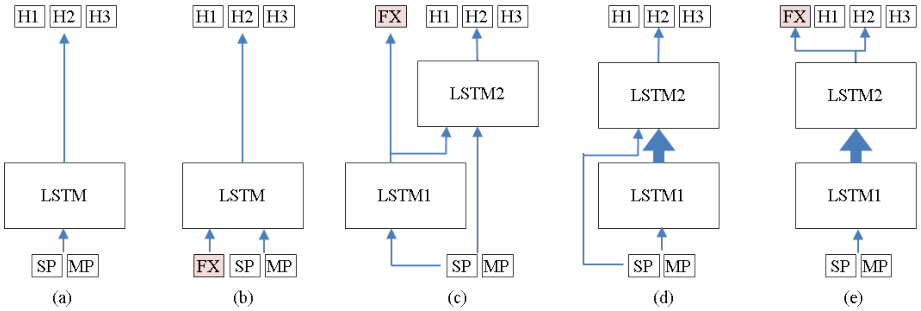


Fig. 3. Single vs. Multi-task models: (a) **Baseline** model with inputs (SP,MP); (b) The **control** model with additional FX modality; (c) **Cascaded** model that combines the prediction of FX by LSTM1 with the prediction of head motions by LSTM2 using combined input; (d) **Cascaded single output** model without the intervening FX-prediction task; (e) **Cascaded multiple outputs** model predicting both FX and Hs by LSTM2

model that is informed by the FX ground truth. Since the correlation of FX with H1 is the highest comparing with H2 and H3, the model is expected to improve significantly the H1 generation quality.

Cascaded. In practice, neither FX nor IU can be used as input feature to train and test data since they are not always available and need to be inferred from observed data such as SP and MP [17]. The incremental estimation of IU is rather difficult with no look-ahead of observations. On the other end, FX are much more likely to be estimated on-line. We thus propose to use a multitask learning, in which FX is generated by a first LSTM layer, called LSTM1 in Figure 3(c). The output of this model is then aggregated with the original input and fed into a second LSTM layer, called LSTM2. This multitask model – with discrete FX and continuous H objectives – is trained in two steps: LSTM1 and LSTM2 are first trained separately and fine-tuning is further performed on the multitask model with both outputs: FX and (H1, H2, H3).

Two other models also have been considered, for fairness:

Cascaded single output. This single-output cascaded model has the same structure as the cascaded model but without the intervening FX-prediction task shown in Figure 3(d).

Cascaded multiple output. Including two LSTMs stacked to each other and predicting both FX and Hs illustrated in Figure 3(e).

5 Results

To compare the performance of each model, leave-one-out cross-validation was performed in which 9 interaction sequences were used to train while the remaining one is used for testing. All models use a total of 80 LSTM neurons. Each

layer of the cascaded models (LSTM1 and LSTM2) has thus 40 LSTM neurons. Pre-training for LSTM1& LSTM2 and fine tuning are both performed with 50 iterations. All models are implemented on Keras with Theano back-end.

Figure 4 displays root mean square error (RMSE) of H1 as a function of number of epochs and model. The control model clearly outperforms the others at epoch 110 with a RMSE of 0.0450 rad. While the *Cascaded* model is able to handle over-fitting and get a minimum RMSE of 0.0569 rad at epoch 76, other methods tend to overfit sooner.

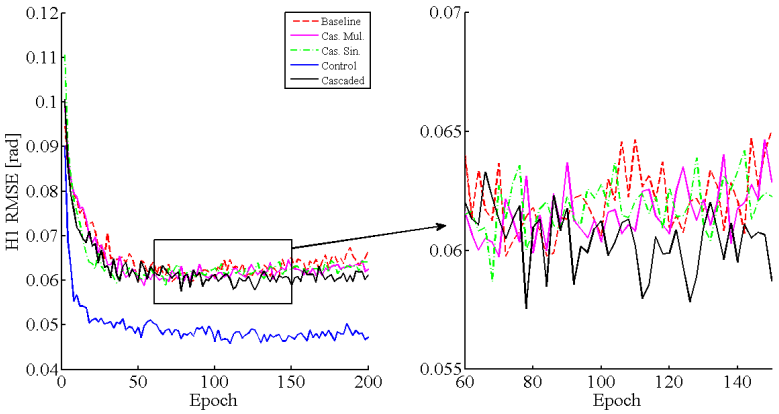


Fig. 4. (Average H1 RMSE at different epochs corresponding to the different cascaded models.

Figure 5 (a) displays a chronogram of ground truth vs. predicted H1. As expected, the *Control* model generates the most faithful movements notably in the vicinity of FX events (see around 7.0 sec). In contrast, the H1 generated by baseline model (driven by the sole SP & MP events) generates delayed head motion. The head motion generated by the cascaded model is close to the one generated by the control model, notably respecting coordination with gaze shifts.

Table 1 gives root mean square errors (RMSE) between ground truth and predicted head motions with different models. As expected from CCA analysis and chronograms, the largest and lowest RMSE are performed respectively by the *Baseline* (0.059) and *Control* (0.045) models. The *Cascaded* model exhibits an intermediate performance (0.057). Since the CCA of FX, SP and MP are not significantly different for H2 and H3, their RMSE are not significantly improved. Pearson correlations, also given in Table 1, corroborate these observations.

In order to compare the micro-coordination patterns, we computed the so-called *coordination histograms* (CH) proposed by Mihoub et al [17]. In order to conform to their proposal, continuous streams of head motions (H1, H2, H3) are first converted to discrete events by detecting peaks of local maximum velocity.

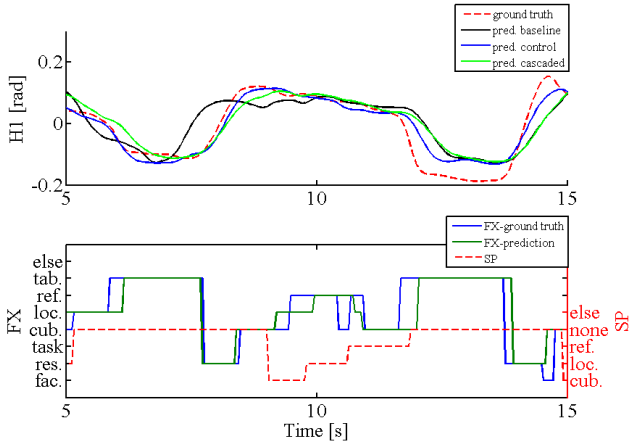


Fig. 5. (a) H1 real vs. prediction streams between different models; (b) input streams (FX ground truth& SP) and FX prediction from LSTM1.

Table 1. Root mean square errors (Pearson correlations) between ground truth and predicted head motions with different models.

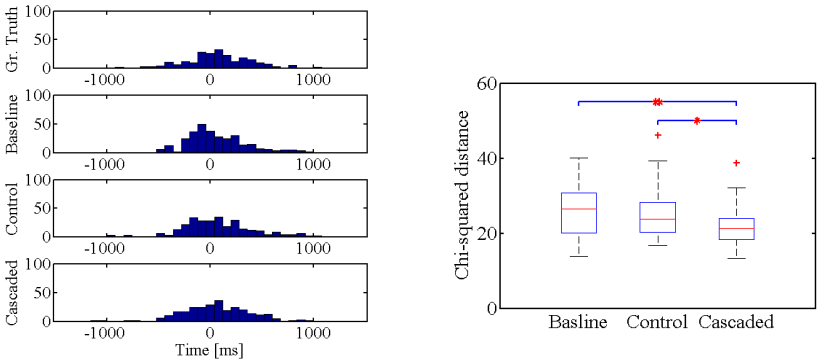
Models	*H1 [rad]*	*H2[rad]*	*H3[rad]*
Baseline	0.059 (0.84)	0.035 (0.67)	0.066 (0.57)
Control	0.045 (0.91)	0.033 (0.72)	0.066 (0.65)
Cascaded	0.057 (0.84)	0.035 (0.67)	0.065 (0.64)

CH are then built by tabulating the time-delay between each event of one given modality and the nearest events from other modalities.

We further compared ground-truth coordinate histogram for H1 with those produced by the different prediction models. Figure 6(a) displays the histogram computed from ground truth vs. CH predicted by the *Baseline*, *Control* vs. *Cascaded* models. Figure 6(b) displays Chi-squared distances between the ground-truth histogram – considered as the target coordination pattern – with those produced by the different prediction models for the three angles. These figures show that the *Cascaded* model outperforms the *Baseline* model both in terms of accuracy and coordination. This result is partly due to the fact that both *Baseline* and *Control* models are directly driven by triggered events (SP, MP) or FX, while the *Cascaded model* is able to handle the coordination pattern – in particular causal relations – between these events.

6 Discussion

Our experiments evidence the benefit that prediction models can draw from a priori knowledge. While recurrent neural networks can implicitly construct



(a) Coordination histograms among H1 and (IU,SP).

(b) Chi-squared distances of the different prediction models.

Fig. 6. Comparing coordination histograms between H1 and (IU,SP) for *Ground truth*, *Baseline*, *Control* vs. *Cascaded* models.

latent representations using massive data, explicit knowledge given as goals or cost functions help them to build and structure intermediate layered mappings.

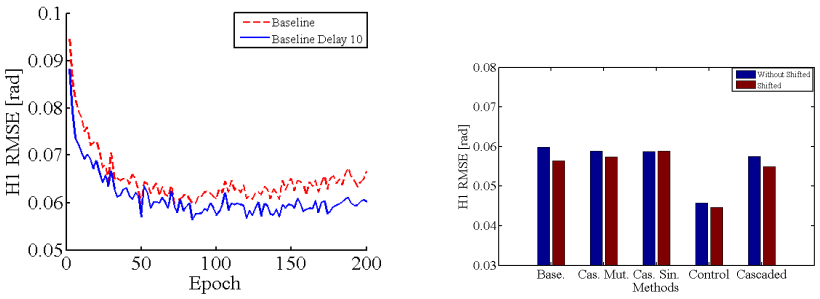
Several algorithms to explore the intra- and inter-slice causal relations between observations have been built for DBN and other statistical models. These analysis tools that help to shape probabilistic graphical models (PGM) may be used to automatically structure neural network layers and give ways to shape latent representations with task-related semantic or pragmatic information.

Although LSTM model can remember events triggered by gaze so that they can drive head motion following the gaze event, it is difficult to deal with subsequent speech events, which are conversely triggered by preceding eye fixations. A way to improve the head motion generation can be done with time shifted speech frames from future, which then becomes the forerunner of head motion. Figure 7(a) displays the H1 RMSE of the baseline model with and without time-shifted input SP frames. SP is here shifted by 10 frames (~ 0.4 sec): it generates lower RMS compared with the original model. Figure 7(b) shows the H1 RMSE obtained at the optimal epoch corresponding to the different models. Almost all shifted SP generate head motion with lower error.

Of course, bidirectional LSTM can be used and combined with a soft attention mechanism to optimally probe contextual information (exogenous as well as intentional). But we here consider reactive models that are able to cope with on-line interactive behaviors: the horizon of the contextual information does not extend beyond the current frame.

7 Conclusions & perspectives

In this paper, we propose an efficient solution to structure the intermediate representations built by layered LSTM. We have shown that gaze can be used



(a) Average H1 RMSE of the *Baseline* model without and with SP shifted frames corresponding to number of training epoch

(b) Average H1 RMSE without and with SP shifted frames corresponding to the different models.

Fig. 7. Comparing H1 RMSE between shifted vs. without time-shifted frames.

effectively as a driving signal for head motion generation. This intervention is effective both in terms of accuracy and coordination patterning.

The quality of prediction may be enhanced in several ways. Other contextual information can be used as additional input – precise regions of interest for the gaze, gaze contacts, communicative functions of speech, etc. – as well as intermediate objectives – e.g. eyebrow movements or respiratory patterns. There, we did not use the segmentation of the task into IUs because most of these IUs were triggered by gaze or speech events. More complex tasks involving switching between multiple interaction styles with multiple agents may motivate the structuring of the interaction by IUs, notably when alternative cues are used to trigger similar pragmatic frames.

Finally, the head motion generation model will be used to drive the head of our iCub-humanoid robot when autonomously instructing human manipulators. We first plan to perform the subjective assessment of our multimodal behavioral model (see [19] for our crowd-sourcing methodology). Another challenge is to adapt this model to multiple manipulators, notably those with motor disabilities. In this case, the behavioral model should both incrementally estimate the best action and the optimal interaction style according to the goodness of fit between the actual and expected behavior of the interlocutor predicted by the joint behavioral model.

8 Acknowledgments

This research is supported by the ANR SOMBRERO (ANR-14-CE27-0014), EQUIPEX ROBOTEX (ANR-10-EQPX-44-01) and the RHUM action of PERSYVAL (11-LABX-0025). The first author is funded by SOMBRERO.

References

1. Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CPVR)*. 961–971.
2. Atef Ben Youssef, Hiroshi Shimodaira, and David A Braude. 2013. Articulatory features for speech-driven head motion synthesis. In *Interspeech*. 2758–2762.
3. Paul Boersma and D Weenik. 1996. PRAAT: a system for doing phonetics by computer. Report of the Institute of Phonetic Sciences of the University of Amsterdam. *Amsterdam: University of Amsterdam* (1996).
4. W Owen Brimijoin, Alan W Boyd, and Michael A Akeroyd. 2013. The contribution of head movement to the externalization and internalization of sounds. *PloS one* 8, 12 (2013), e83068.
5. Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 3 (2007), 1075–1086.
6. Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Annual conference on Computer graphics and interactive techniques*. ACM, 413–420.
7. Catherine Dehon, Peter Filzmoser, and Christophe Croux. 2000. Robust methods for canonical correlation analysis. *Data analysis, classification, and related methods* (2000), 321–326.
8. Yu Ding, Catherine Pelachaud, and Thierry Artieres. 2013. Modeling multimodal behaviors from speech prosody. In *International Workshop on Intelligent Virtual Agents (IVA)*. Springer, 217–228.
9. Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. 2002. Visual prosody: Facial movements accompanying speech. In *Automatic Face and Gesture Recognition (FG)*. IEEE, 396–401.
10. D Guitton and M Volle. 1987. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology* 58, 3 (1987), 427–459.
11. Kathrin Haag and Hiroshi Shimodaira. 2016. Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis. In *International Conference on Intelligent Virtual Agents (IVA)*. Springer, 198–207.
12. Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents (IVA)*. Springer, 243–255.
13. Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. In *ACM Transactions on Graphics (TOG)*, Vol. 28. ACM, 172.
14. Chaoran Liu, Carlos T Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Human-Robot Interaction (HRI)*. IEEE, 285–292.
15. Soroosh Mariooryad and Carlos Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.

16. Tobias May, Ning Ma, and Guy J Brown. 2015. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2679–2683.
17. Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2016. Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters (PRL)* 74 (2016), 82–89.
18. Kevin G Munhall, Jeffery A Jones, Daniel E Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2004. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science* 15, 2 (2004), 133–137.
19. Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. 2016. Conducting neuropsychological tests with a humanoid robot: design and evaluation. In *Cognitive Infocommunications (CogInfoCom)*. IEEE, 337–342.
20. Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. (accepted with minor revision). Learning Off-line vs. On-line Models of Interactive Multimodal Behaviors with Recurrent Neural Networks. *Pattern Recognition Letters (PRL)* ((accepted with minor revision)).
21. Najmeh Sadoughi and Carlos Busso. 2017. Speech-driven Animation with Meaningful Behaviors. *arXiv preprint arXiv:1708.01640* (2017).
22. Kristinn R Thórisson. 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In *Multimodality in language and speech systems*. Springer, 173–207.
23. Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *International Conference on Language Resources and Evaluation (LREC)*.
24. Hani Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. 2000. Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Germany, 265–268.
25. Daniel M. Wolpert, Kenji Doya, and Mitsuo Kawato, 2003. A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358, 1431 (2003), 593–602.