# Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions

*Bekir Berker Türker, Engin Erzin, Yücel Yemez, Metin Sezgin*

Koç University, Turkey

bturker13,eerzin,yyemez,mtsezgin@ku.edu.tr

## Abstract

Head-nods and turn-taking both significantly contribute conversational dynamics in dyadic interactions. Timely prediction and use of these events is quite valuable for dialog management systems in human-robot interaction. In this study, we present an audio-visual prediction framework for the head-nod and turn-taking events that can also be utilized in real-time systems. Prediction systems based on Support Vector Machines (SVM) and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) are trained on human-human conversational data. Unimodal and multi-modal classification performances of head-nod and turn-taking events are reported over the IEMOCAP dataset.
**Index Terms**: head-nod, turn-taking, social signals, event prediction, dyadic conversations, human-robot interaction.

## 1. Introduction

Recent intelligent human-computer interaction (HCI) literature widely includes studies on language, dialogue management and speech recognition [1]. Targeting a more natural, flexible and generalizable HCI still sets challenging problems. Since early 2000s, inception of new research fields, such as investigation of non-verbal cues for human-human interaction, has enabled technologies for more humane (human-like) HCI systems [2]. Robots and virtual agents are expected to understand what the user says, as well as to monitor the users emotional and/or cognitive state and their audio/visual reactions (gestures, views, facial expressions, mimics etc.), to appropriately take more humane actions in the course of HCI. In this way, it is expected that the intermediaries will be more convincing and natural, and that they will keep the user engaged and enable them to interact more efficiently [3, 4, 5, 6, 7, 8, 9, 10].

In this study, we focus on head-nod and turn-taking events that are quite functional in human-human and human-robot interactions. As humans, we manage the conversational flow with smooth turn-takings and execute timely head-nods for emphasis or feedback. On the other hand, it is a challenging task to predict timely turn-taking or head-nod events to improve naturalness and user engagement in human-robot interactions. Furthermore, these two events help monitoring and sustaining the user engagement [11].

We construct a multi-modal framework for prediction of the head-nod and turn-taking events in dyadic conversations. The prediction task is basically a binary decision for a particular event which is likely to happen in the upcoming time instant. Hence, the problem can be defined as observing dyadic signals in time interval $[t-c, t]$ to make a prediction at time $t$ for the starting event at time $t+d$, where $c$ is the duration of the temporal window and $d$ is the time till the event start.

The rest of the paper is organized as follows. Section 2 presents literature review. Section 3 defines the event prediction framework for dyadic interaction setup. Section 4 presents experimental work of the proposed head-nod and turn-taking event prediction framework, and finally Section 5 gives the conclusion.

## 2. Related Work

In the literature, head-nod recognition and detection have been studied extensively [12, 13, 14]. However, head-nod timing prediction has been addressed in fewer number of studies.

In some of the works, head-nod is addressed under prediction of backchannels which are shortly defined as non-intrusive feedback expressions [15]. The existing backchannel feedback mechanisms used in the human-computer interaction are often founded on rule-based approaches using simple statistical data [4, 16, 17, 18, 19]. For instance, the relation between the changes in the sound perception of the speaker and the triggered backchannel signals of the listener has been examined in [16]. In this study, a backchannel signal is triggered when there are pauses with certain lengths, preceded by an increase or decrease in the sound pitch. In this estimation approach, the type of backchannel signal is not taken into consideration. In the SAL (Sensitive Artificial Listeners) system [4], the engagement level and emotional response of the user is monitored, and events where backchannel feedback is necessary are observed. For example, when the user shakes their head or there is a change in their sound pitch, it is understood that a feedback is necessary. According to the current mood state of the agent, through a predefined decision process a backchannel feedback such as smile or head nod/shake is synthesized.

There are very few studies in the literature aiming to learn backchannel by using a model based learning [20, 21, 22]. For example, the head shake gesture as a backchannel is estimated from speech prosody, spoken words, and eye movements using Hidden Markov Models in [20]. Their estimation results are compared with reference data, but the proposed method is not implemented under any human-computer interaction scenario. Whereas in [21], only the verbal backchannel signals are predicted based on linguistic and prosodic cues using the naive Bayesian classifier. This classifier is tested on a verbal interface with a dialogue management mechanism, and it is reported to yield better results than an approach that randomly generates a backchannel signal. Again [22], predict the head nod/shake by hidden Markov models in relation to verbal and emotional features of the interaction.

Turn-taking prediction is studied by many research groups since organization of the turns has been a problem in human-robot/agent interactions. Throughout the studies and observations in human-human interactions, prosodic changes [23, 24] and eye-gaze patterns [25, 23] are shown that they are leading factors in turn-taking behaviours. Kawahara et al. [23], combine para-linguistic and non-verbal patterns to detect speaker change and to determine next speaker in poster sessions. They

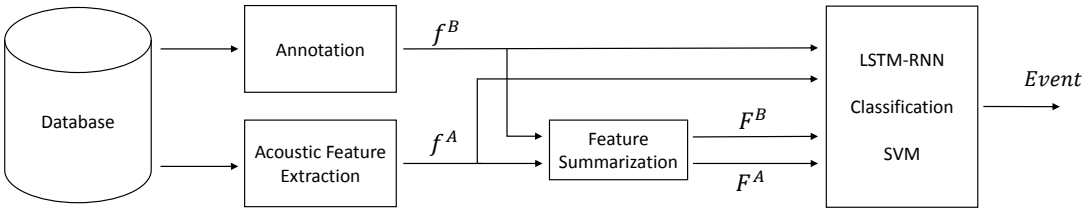


Figure 1: *System overview*

state that prosodic and gaze features are useful for speaker change detection, while backchannel of audience is also useful to determine next speaker. In a more recent study, Skantze [24] proposed a continuous model solution for turn-taking problems. His work is solely based on audio features where visual channel communication is not considered. He used task-based interaction database in training while we use naturalistic social talk interactions which is more challenging and applicable in wild. Our work also contributes in prediction of head-nod which is a social signal and supporting turn-taking predictions.

# 3. Methodology

Our main objective in this study is to predict head-nod and turn-taking events before they take place in a dyadic conversation setup. We value these two problems for a more natural HCI system, in which a robot or virtual agent can predict probable head-nod or turn-taking opportunity by monitoring the interaction between human user and itself. In this purpose, we propose an event prediction framework, which can be trained using human-human dyadic conversational data. Figure 1 presents system overview of the proposed framework. There are two sets of features which are extracted from low-level acoustic signals and high-level non-verbal behavioral cues. We perform feature summarization for the SVM classifier and utilize temporal feature stream for the LSTM-RNN classifier. Each block of the framework and their input/output characterizations are given in the following subsections.

We define turn-taking event as the time instant that user ends her/his turn. Similarly, head-nod event is the time instant that head-nod action starts. Regardless of their type, we call them both '*event*' which occurs at given time $t$. Figure 2 shows event prediction structure over the interaction time-line of participants 1 and 2. Binary decision of the event occurrence is predicted by using features over the recent temporal window of length $c$ seconds.

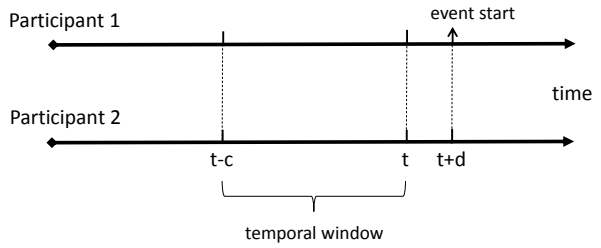


Figure 2: *Event prediction time-line*

We define two sets of features for the event prediction over the temporal window. Section 3.1 defines the first set as low-level acoustic features. The second set of features are defined on high-level non-verbal behavioral cues in Section 3.2. In the following subsections, aforementioned features and corresponding dimensions are provided for single participant. Eventual feature dimensions are given in Section 4, since minor changes apply for different event predictions: turn-taking and head-nod.

### 3.1. Acoustic Features

Acoustic features are composed of both spectral and prosodic features. Spectral properties are described by mel-frequency cepstral coefficient (MFCC) representation which is the most commonly used spectral feature in speech and audio processing. We compute 12-dimensional MFCC features. Also, the log-energy coefficient is appended. Thus, the resulting 13-dimensional spectral feature vector is defined as $f^M$.

Prosody characteristics at the acoustic level carry important temporal and structural clues for audio portion just before the event occurrence. We choose to include speech intensity, pitch, and confidence-to-pitch into the prosody feature vector as in [26, 27]. Speech intensity is extracted as the logarithm of the average signal energy over the analysis window. Pitch is extracted using a well-known auto-correlation based method [28]. Confidence-to-pitch provides an auto-correlation score for the fundamental frequency [27]. These three parameters and their first temporal derivatives form the 6-dimensional prosody feature vector, denoted by $f^P$.

Since extracted acoustic features are speaker and utterance dependent, we apply mean and variance normalization to both spectral and prosodic features. The mean and variance normalization of features is performed over the temporal window. Then the normalized spectral and prosodic features are concatenated and resulting 19-dimensional acoustic feature vector is obtained: $f^A = [f^M f^P]$. Note that acoustic features are extracted using a 40 msec sliding window at intervals of 25 msec. Thus, feature frame rate is 40 Hz.

### 3.2. Non-verbal behavioral cues

Non-verbal behavioral cues are produced by participants in the interactions, mostly in a unconscious way. These cues create a harmony between the participants, for instance one participant's head-nod may trigger the other's head-nod. Mirroring is defined as unconsciously copying others' non-verbal expressions in an interaction [29]. Similarly, mirroring effect is observed in laugh/smile expressions [30]. On the other hand, gazing behaviours have important roles in social interactions. In [31], authors show eye-gaze convey essential cues for turn-taking. Also, turn-taking request can be expressed with relatively rapid, large and frequent head-nodding [32].

Considering the meanings and their interactions of these non-verbal behavioral cues, we choose to include them into the feature set of our event prediction framework. The labels over the time-line is transformed into binary values with frame rate

of 40 Hz. The frames get 1's where non-verbal behavioral event is active, the rest of the regions get 0's which means the event is not active. The frames are totally synchronous with acoustic feature frames. The resulting 3-dimensional non-verbal behavioral cues (head-nod, laugh/smile, gaze away) feature vector is denoted as $f^B$.

### 3.3. Feature Summarization

We perform statistical feature summarization over the temporal window $[t-c, t]$ for the SVM classifier. For this purpose, we compute statistical quantities of the acoustic feature that comprise of 11 functionals, which are the mean, standard deviation, skewness, kurtosis, range, minimum, maximum, first quantile, third quantile, median quantile and inter-quantile range. This set of functionals were successfully used before by Metallinou et al. [33] for continuous emotion recognition from speech and body motion. Resulting summarized acoustic feature vector is denoted as $F^A$. The dimension of $F^A$ is 11 times the dimension of $f^A$.

Similarly, non-verbal behavioral cues features are reduced down to single vector. Since these features are binary values, we choose to keep only one value which denotes the existence indication of the non-verbal behavioral cue in a given temporal window, $[t-c, t]$. For instance, if any portion of head-nod event appears within the temporal window, we summarize it as 1, otherwise as 0. Thus, summarized single vector has 3 dimensions and denoted as $F^B$.

### 3.4. Classifier

We employ SVM and LSTM-RNN classifiers in the proposed event prediction framework. SVM is a binary classifier based on statistical learning theory, which maximizes the margin that separates samples from two classes [34]. SVM projects data samples to different spaces through kernels that range from simple linear to radial basis function (RBF) [35]. We consider the summarized statistical features ($F^A$ and $F^B$) as inputs of the SVM classifier to discriminate occurance or absence of the 'event' at time $t+d$.

On the other hand, LSTM-RNN is an recurrent artificial neural network model. We use the LSTM-RNN classifier to model the temporal structure of the feature streams as it has been successfully used in many time-series data [36]. Since the model is sequential, frame based time-series features ($f^A$ and $f^B$) are used without summarization. In this work, we use only one LSTM layer. The further details of the classifier structure and its parameters are explained in Section 4.

## 4. Experiments and Results

In Section 3, we describe the methodology of event prediction in dyadic interactions. Following that, we perform prediction of two events: turn-taking and head-nod. During the experiments, the person of interest (POI) is either Participant 1 or Participant 2. In other words, we have a single classification structure which takes one participant on focus and make prediction for that participant. For instance, each annotated head-nod belongs to one participant position: either 1 or 2. Thus, it is known which participant is POI for prediction of each head-nod. Then, we have 2 sources of features: POI and the other (OTH). Naturally, acoustic features and non-verbal behavioral cues features are available for both of them.

In turn-taking prediction, we use POI($f^A$, $f^B$) and OTH($f^B$). Since, we know only POI has speech in temporal window. On the other hand, in head-nod prediction we use all available features: POI($f^A$, $f^B$) and OTH($f^A$, $f^B$). Note that feature dimension orders are preserved both in between POI, OTH and $f^A$, $f^B$.

### 4.1. Dataset and Annotations

In order to carry out experimental work, IEMOCAP database [37] is used. IEMOCAP consists of naturalistic human-human dyadic conversations with rich affective contents. Five dyads (sessions), interacts under scripted and spontaneous scenarios, resulting with 8 hours of data.

Annotations of non-verbal behavioural cues are carried out on IEMOCAP database. Laugh/smile annotation performed in our previous study [38]. Head-nod and gaze away cues are annotated by two human subjects. Then, the annotations are checked and if necessary corrected by a third subject.

In our annotation scheme, head-nod is defined as vertical swing of head for a reasonable of time with a conscious or unconscious mission of carrying a message to other interlocutor. This definition helped eliminating highly speaker dependent head-nod behaviours. On the other hand, gaze away is defined as the time portions that the gaze has a fixation out of the other interlocutor. Rapid eye movements (saccades) are not annotated as gaze away.

Table 1: *Annotation statistics of the three non-verbal behavioral cues and turn segments that are longer than 5 seconds*

| | **# of events** | **Duration (sec)** | |
|---|---|---|---|
| | | **Mean** | **Std** |
| **Head-Nod** | 1648 | 1.25 | 0.67 |
| **Laugh/smile** | 1244 | 0.96 | 0.91 |
| **Gaze Away** | 5147 | 4.26 | 5.14 |
| **Turns** | 3132 | 8.04 | 3.02 |

Table 1 reports basic annotation statistics of the three non-verbal behavioral cues and turn segments that are longer than 5 seconds. Turn annotations comes with IEMOCAP database as described in [37].

### 4.2. Training

We set the temporal window size $c = 2$ sec for turn-taking and $c = 3$ sec for head-nod event prediction. Then, we create a class balanced dataset for each events. For head-nod events, we obtain balanced dataset by randomly picking 1648 negative class samples from no head-nod regions. For turn-taking events, for each turn region, we extract one positive $[t-2, t]$ and one negative sample $[t-4, t-2]$. We keep only turns longer than 5 sec since we experiment with varying $d = \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. We attain balanced dataset with 6264 samples. In all experiments, these balanced datasets are used in one-session-out (5-fold) test fashion. Thus, all results are speaker independent.

RBF kernel is used in SVM training and hyper-parameters are optimized in the first fold with grid search over cross-validation scheme. Optimized parameters are kept fixed for the rest folds.

In training of LSTM-RNN, we used single layer of LSTM with 50 hidden nodes. During the 5-fold test, in each fold we split 1 session for validation and train with 3 sessions. Early-stopping is provided by the best accuracy performance on validation set.

### 4.3. Results

In this section, we report our experimental evaluations, which are obtained over the balanced datasets created in Section 4.2. Since, experiments are over class balanced set, widely used performance metrics, accuracy (Acc), precision (Pre), recall (Rec), $F_1 - score$ ($F_1$), are utilized in the evaluations. Note that, positive class is happening of the event: head-nodding of POI, turn-ending of POI (turn-taking of OTH).

Table 2: *Turn-taking prediction performances with the SVM for varying d in seconds*

| $d$ | **0** | **0.1** | **0.2** | **0.4** | **0.6** | **0.8** | **1** |
|---|---|---|---|---|---|---|---|
| **Acc** | 75,69 | 73,65 | 71,57 | 65,72 | 58,96 | 57,27 | 54,92 |
| **Pre** | 74,98 | 72,79 | 70,71 | 64,92 | 58,13 | 56,36 | 54,38 |
| **Rec** | 77,10 | 75,56 | 73,65 | 68,40 | 64,11 | 64,43 | 61,11 |
| $F_1$ | 76,03 | 74,15 | 72,15 | 66,61 | 60,97 | 60,13 | 57,55 |

Turn-taking prediction performances with SVM classifier are given in Table 2. The best performances are observed with the minimum delay till the event, $d = 0$, which is expected by the nature of the problem. We expect the largest differentiation between positive and negative class samples when $d = 0$ and thus better performance than the others, $d > 0$.

Considering human-robot interaction, giving 200 msec advance to the robot to take the turn could be very beneficial even though prediction performances degrade slightly at $d = 0.2$ sec. Another promising observation is having close precision and recall performances since both are important for the task. High precision means less false positives and thus less intervention while user's turn continues. High recall means less false negatives and thus less lagging to take the turn.

Table 3: *Turn-taking prediction performances with the LSTM-RNN for varying d in seconds*

| $d$ | **0** | **0.1** | **0.2** | **0.4** | **0.6** | **0.8** | **1** |
|---|---|---|---|---|---|---|---|
| **Acc** | 83,62 | 81,91 | 79,89 | 72,94 | 64,45 | 61,14 | 58,42 |
| **Pre** | 80,53 | 78,21 | 77,51 | 71,59 | 63,47 | 61,64 | 57,98 |
| **Rec** | 88,68 | 88,46 | 84,23 | 76,07 | 68,07 | 58,98 | 61,14 |
| $F_1$ | 84,41 | 83,02 | 80,73 | 73,76 | 65,69 | 60,28 | 59,52 |

Table 3 presents turn-taking prediction performances using LSTM-RNN. We observe remarkable performance improvement with LSTM-RNN compared to SVM for small $d$. As $d$ increases, performance difference is getting smaller, which is quite legitimate since negative and positive class samples getting so similar and makes hard to capture pattern differences for both of the classifiers. Another observation is that LSTM-RNN produces systems with better recall performances. This can be useful for a robot, which displays more talking with more generous interventions but also timely turn-takings.

Table 4: *Head-nod prediction performances with SVM and LSTM-RNN*

| | **SVM** | | **LSTM-RNN** | |
|---|---|---|---|---|
| | $F^A, F^B$ | $F^A, F^B, F^{SA}$ | $f^A, f^B$ | $f^A, f^B, f^{SA}$ |
| **Acc** | 59,46 | 62,49 | 59,01 | 63,37 |
| **Pre** | 60,10 | 63,53 | 61,75 | 64,21 |
| **Rec** | 56,13 | 58,56 | 47,21 | 60,32 |
| $F_1$ | 58,05 | 60,94 | 53,51 | 62,20 |

In head-nod prediction task, we consider speech activity (turn states) to help prediction of head-nods, as turn-taking prediction already owns head-nod features. Thus we extract speech activity feature $f^{SA}$, which is binary valued stream obtained from turn annotations as defined for the other non-verbal behavioral cues features. Similarly, summarized speech activity feature $F^{SA}$ is just defined as a binary value to indicate any temporal change in $f^{SA}$. The $F^{SA}$ indicates if the participant started talking or stopped talking.

In Table 4, head-nod prediction performances are given with and without the speech activity features. We observe that speech activity features improve performances for both classifiers, but LSTM-RNN favors more than SVM. Although unimodal head-nod prediction performances of acoustic features and social cues are poor and close to random, the multimodal performances are promising. Since precision is higher, it indicates less false positives and thus less awkward timely head-nods. False negatives do not create much trouble, since passing head-nod opportunity as a robot is not a big problem. Actually, even humans have much variety on head-nodding behaviour. One can just pass head-nodding if she/he experienced very same moment that she/he head-nodded before.

We perform decision fusion of the best performing SVM and LSTM classifiers by weighted sum of the classifier confidence scores as, $\alpha * SVM(F^A, F^B, F^{SA}) + (1 - \alpha) * LSTM(f^A, f^B, f^{SA})$, where $\alpha$ is the weight of the SVM confidence score. Table 5 presents head-nod prediction performances of the decision fusion for three values of $\alpha$. We observed that the decision fusion with $\alpha = 0.6$ produced the highest $F_1 - scores$ for head-nod prediction task.

Table 5: *Decision fusion of the best performing SVM and LSTM classifiers: SVM($F^A, F^B, F^{SA}$) and LSTM($f^A, f^B, f^{SA}$)*

| $\alpha$ | **0.5** | **0.6** | **0.7** |
|---|---|---|---|
| **Acc** | 64,28 | 65,28 | 65,34 |
| **Pre** | 65,56 | 66,80 | 67,17 |
| **Rec** | 60,07 | 60,68 | 59,95 |
| $F_1$ | 62,70 | 63,59 | 63,35 |

## 5. Conclusion

In this paper, we present a generalized framework for event prediction in dyadic interactions, such as human-human and human-robot. We report turn-taking and head-nod prediction performances over human-human conversational data. We showed that turn-taking prediction can be achieved with relatively better performance, even for predictions hundreds of milliseconds ahead. Head-nod prediction experiments showed that speech activity features have potential to improve the performance and fusion of classifiers achieves the best overall performance.

The proposed framework has a potential use for more humane human-robot interactions. Smooth turn-takings and producing timely head-nods are expected to make robots more natural and engaging.

## 6. Acknowledgements

This work is supported by Turkish Scientific and Technical Research Council (TUBITAK) under grant numbers 113E324 and 217E040.

# 7. References


[1] V. Zue and J. Glass, "Conversational interfaces: Advances and challenges," *Proc. IEEE*, vol. 42, pp. 1166–1180, 2000.

[2] C. Clavel, A. Cafaro, S. Campano, and C. Pelachaud, *Fostering User Engagement in Face-to-Face Human-Agent Interactions: A Survey*. Cham: Springer International Publishing, 2016, pp. 93–120.

[3] S. DMello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *ACM Trans Interact Intell Syst*, vol. 4, no. 2, pp. 1–39, 2013.

[4] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, 2012.

[5] A. Choi, C. D. Melo, W. Woo, and J. Gratch, "Affective engagement to emotional facial expressions of embodied social agents in a decision-making game," *Computer Animation and Virtual Worlds*, vol. 23, no. 3-4, pp. 331–342, 2012.

[6] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki, "Activating humans with humor–a dialogue system that users want to interact with," *IEICE transactions on information and systems*, vol. 92, no. 12, pp. 2394–2401, 2009.

[7] T. Bickmore, L. Pfeifer, and D. Schulman, "Relational agents improve engagement and learning in science museum visitors," in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 55–67.

[8] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proceedings of the 16th international conference on multimodal interaction*. ACM, 2014, pp. 2–9.

[9] E. Andre, M. Rehm, W. Minker, and D. Bühler, "Endowing spoken language dialogue systems with emotional intelligence," in *Tutorial and Research Workshop on Affective Dialogue Systems*. Springer, 2004, pp. 178–187.

[10] M. Ptaszynski, P. Dybala, R. Rzepka, and K. Araki, "Forgetful and emotional: Recent progress in development of dynamic memory management system for conversational agents," in *Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium*, 2010, pp. 32–38.

[11] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2010, pp. 375–382.

[12] J. Zhao and R. S. Allison, "Real-time head gesture recognition on head-mounted displays using cascaded hidden markov models," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2017, pp. 2361–2366.

[13] A. Kapoor and R. W. Picard, "A real-time head nod and shake detector," in *Proceedings of the 2001 workshop on Perceptive user interfaces*. ACM, 2001, pp. 1–5.

[14] Y. Chen, Y. Yu, and J.-M. Odobez, "Head nod detection from a full 3d model," in *Proceedings of the ICCV 2015*, no. EPFL-CONF-213704, 2015.

[15] V. H. Yngve, "On getting a word in edgewise," in *Chicago Linguistics Society, 6th Meeting*, 1970, pp. 567–578.

[16] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), September 2010, pp. 3058–3061.

[17] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth, "Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 181–188.

[18] R. Poppe, K. P. Truong, and D. Heylen, "Perceptual evaluation of backchannel strategies for artificial listeners," *Autonomous Agents and Multi-Agent Systems*, vol. 27, no. 2, pp. 235–253, 2013.

[19] R. Poppe, M. ter Maat, and D. Heylen, "Switching wizard of oz for the online evaluation of backchannel behavior," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 109–117, 2014.

[20] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10458-009-9092-y

[21] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech & Language*, vol. 28, no. 4, pp. 903–922, 2014.

[22] J. Lee and S. C. Marsella, "Predicting speaker head nods and the effects of affective information," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 552–562, 2010.

[23] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[24] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 220–230.

[25] D. G. Novick, B. Hansen, and K. Ward, "Coordinating turn-taking with gaze," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, Oct 1996, pp. 1888–1891 vol.3.

[26] E. Bozkurt, E. Erzin, and Y. Yemez, "Affect-Expressive Hand Gestures Synthesis and Animation," in *IEEE International Conference on Multimedia and Expo (ICME)*, Torino, Italy, 2015.

[27] E. Bozkurt, Y. Yemez, and E. Erzin, "Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures," *Speech Communication*, vol. 85, pp. 29–42, 2016.

[28] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, p. 1917, 2002.

[29] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction." *Journal of personality and social psychology*, vol. 76, no. 6, p. 893, 1999.

[30] S. Finger, "Curious behavior: Yawning, laughing, hiccupping, and beyond by robert provine," *Journal of the History of the Neurosciences*, vol. 22, no. 4, pp. 429–430, 2013.

[31] S. Ho, T. Foulsham, and A. Kingstone, "Speaking and listening with the eyes: gaze signaling during dyadic interactions," *PloS one*, vol. 10, no. 8, p. e0136905, 2015.

[32] J. Napier and L. Leeson, "Sign language in action," in *Sign Language in Action*. Springer, 2016, pp. 50–84.

[33] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[35] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[38] B. B. Turker, Y. Yemez, T. M. Sezgin, and E. Erzin, "Audio-facial laughter detection in naturalistic dyadic conversations," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 534–545, Oct 2017.