Prediction of Visual Backchannels in the Absence of Visual Context Using Mutual Influence

Derya Ozkan and Louis-Philippe Morency

University of Southern Caligornia
Institute for Creative Technologies
10215 Waterfront Drive, Playa Vista, CA
{ozkan,morency}@ict.usc.edu

Abstract. Based on the phenomena of mutual influence between participants of a face-to-face conversation, we propose a context-based prediction approach for modeling visual backchannels. Our goal is to create intelligent virtual listeners with the ability of providing backchannel feedbacks, enabling natural and fluid interactions. In our proposed approach, we first anticipate the speaker behaviors, and then use this anticipated visual context to obtain more accurate listener backchannel moments. We model the mutual influence between speaker and listener gestures using a latent variable sequential model. We compared our approach with state-of-the-art prediction models on a publicly available dataset and showed importance of modeling the mutual influence between the speaker and the listener.

Keywords: nonverbal behavior, embodied conversational agent.

1 Introduction

During face-to-face communication, participants often mutually influence each other through their verbal and nonverbal behaviors. For instance, a speaker will decide to give more explanations or simply continue with the story based on the feedbacks from the listener. Similarly, participants often mimic each others gestures to convey empathy and rapport [1–3]. This phenomena, which we refer as mutual influence in this paper, is essential for fluid human interactions; but research is still needed to replicate this process with virtual humans.

A good example of human behaviors that involves mutual influence is backchannel feedbacks (i.e. the nods and paraverbal signals such as "uh-hu" and "mm-hmm" that listeners produce as someone is speaking). Backchannel feedbacks have received considerable attention due to their pervasiveness across languages and conversational contexts. They play a significant role in determining the nature of a social exchange by showing rapport and engagement [4]. When these signals are positive, coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in diverse areas such as negotiations and conflict resolution [5], psychotherapeutic effectiveness [6], improved test performance

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 189–202, 2013.

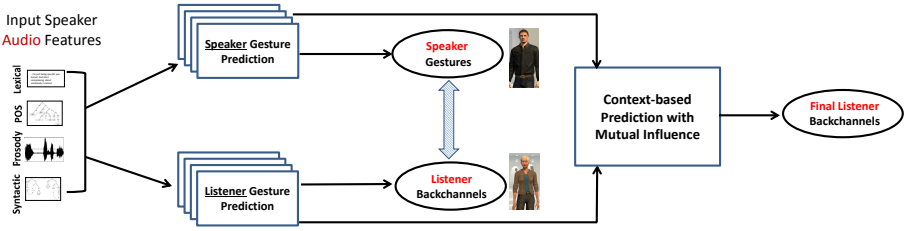

Fig. 1. An overview of our approach for predicting listener backchannels in the absence of visual information. Our approach takes into account the context from the speaker by first predicting the nonverbal behaviors of the speaker and uses these predictions to improve the final listener backchannels.

in classrooms [7] and improved quality of child care [8]. By correctly predicting backchannel feedback, we can improve the way that the machines communicate with human. For instance, a virtual human that provides head nods at reasonable points in the conversation can have a stronger sense of rapport.

One of the challenges in building intelligent virtual agents with such abilities is absence of the visual information. In many real-world applications, we often have only the speech and/or text to be spoken by the virtual human, without any visual context. Another scenario where no visual context is available is phone-to-phone conversations. If we want to create a virtual (i.e. customer service) representative that is capable of providing backchannel feedbacks, the only source of information is the interlocutor's (customer's) voice. As discussed above, a good prediction model of backchannels should be able to take into account the mutual influence between participants even in the absence of visual context. This can be achieved by anticipating the nonverbal behaviors of the speaker, and using the anticipated visual context to model the mutual influence between the speaker and the listener.

In this paper, we present a context-based prediction model to predict backchannels of a listener during dyadic conversations. An overview of our approach is given in Figure 1. We assume an environment where the visual gestures of the speaker are not available. Based on this assumption, we first predict the visual context (i.e. nonverbal behaviors) of the speaker and the backchannels of the listener using only the auditory observations (features) from the speaker. We model the mutual influence between the speaker and the listener by using a latent variable model based on Latent Mixture of Discriminative Experts (LMDE) [9]. We evaluate our approach using 45 storytelling dyadic interactions from the RAPPORT dataset [10]. In our experiments, we compare our approach with previous approaches based on Conditional Random Fields (CRF) [11], Latent-Dynamic CRFs [12], and CRF Mixture of Experts (a.k.a Logarithmic Opinion Pools [13]), and a rule based random predictor [14].

The paper is organized as follows. We first present the related works in Section 2. Then we present our context-based prediction approach in Section 3.

Experimental setup, and results are given in Section 4 and Section 5, respectively. Finally, we conclude in Section 6.

2 Related Works

Although human can naturally display and interpret nonverbal signals in social context, computers are not equipped with such abilities. Therefore, supporting such fluid interactions has become an important topic in computer science research [15]. Many different models have been proposed to recognize [16, 17], or predict [14, 18, 10] certain nonverbal behaviors.

The application described in this paper uses audio cues from the speaker to predict the social behavior of the participant. This type of predictive models has been mostly studied in the context of embodied conversational agents [19, 20]. Several researchers have developed models to predict when backchannel should happen. In general, these results are difficult to compare as they utilize different corpora and present varying evaluation metrics. Ward and Tsukahara [14] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Later in 2003, Ward [21] studied both the forms and functions of sounds like h-nmm, hh-aaaah, hn-hn, unkay, nyeah, ummum, uuh and um-hmuh -hm in American English conversation.

Fujie et al. [22] use Hidden Markov Models to perform head nod recognition. In their proposal, they combined head gesture detection with prosodic low-level features from the same person to determine strongly positive, weak positive and negative responses to yes/no type utterances. Maatman et al. [18] present a multimodal approach where Ward and Tsukhara's prosodic algorithm is combined with a simple method of mimicking head nods. No formal evaluation of the predictive accuracy of the approach was provided but subsequent evaluations have demonstrated that generated behaviors do improve subjective feelings of rapport [23] and speech fluency [4]. Morency et al. [10] showed that Conditional Random Field models can be used to learn predictive features of backchannel feedback. In their approach, multimodal features are simply concatenated in one large feature vector for the CRF model. They show statistical improvement when compared to the rule-based approach of Ward and Tsukahara [14].

The Semaine Project of EU-FP7 [24] focuses on building *Sensitive Artificial Listeners*. Towards this effort, Gravano [25] focuses on backchannel-inviting cues as part of as part of their study of turn-taking phenomena. They first analyze individual acoustic, prosodic and textual backchannel-inviting cues; then, they investigate how such cues combine together to form complex signals. In [26], Neiber focuses on the communicative functions of vocal feedback like "mhm", "okay" and "yeah, thats right". They categorize feedback as non-lexical, lexical and phrase based feedback.

In this paper, we present an approach to predict the backchannels of a listener using the anticipated visual context of the speaker. More specifically, we focus on the visual feedbacks of the listener: head nods. We assume an environment, in which the visual information for both the listener and the speaker is absent.

3 Context-based Backchannel Prediction with Mutual Influence

The goal of our approach is to predict listener backchannels in dyadic conversations by using the mutual influence between the speaker and the listener. We assume a situation where no visual context from neither the speaker nor the listener is available. In other words, we have no access to speaker's visual information, but only the speech/text information from the speaker. In our approach, we explicitly model multiple dimensions of the speech information such as prosody, lexicons, syntactic structure and part-of-speech tags. These different dimensions contain complementary information, and our approach will model the hidden dynamic between them.

In our context-based prediction approach, we first infer the speaker gestures, and then exploit this visual context to improve the final listener backchannel predictions (see Figure 1). In order to model the mutual influence between the speaker and the listener, we use a variant of the Latent Mixture of Discriminative Experts

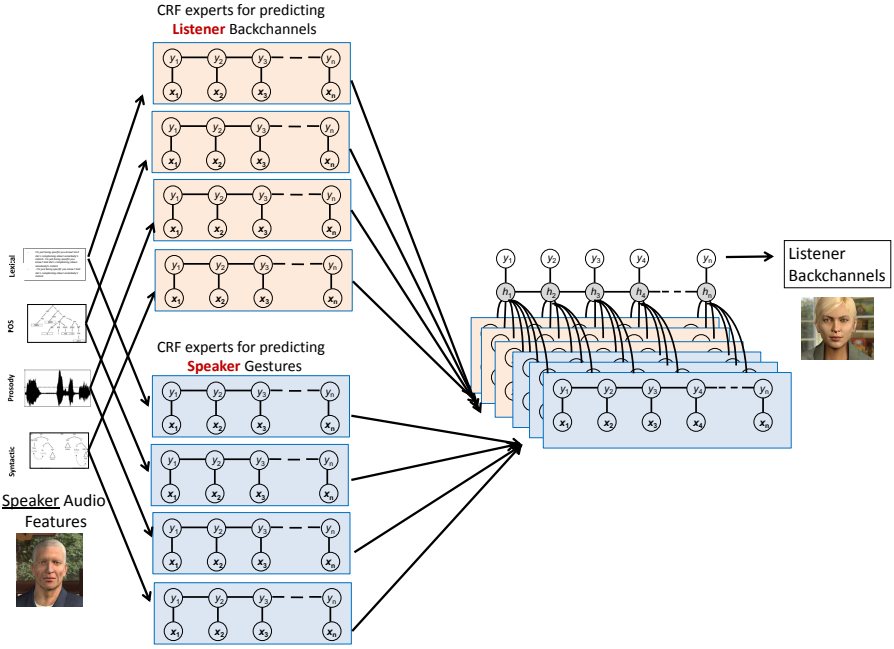

Fig. 2. Our approach for predicting speaker gestures in dyadic conversations. Using the speaker audio features as input, we first learn a CRF model (expert) per each audio channel and for both speaker gestures and listener backchannels. Then, we merge these CRF experts using a latent variable model that is capable of learning the hidden dynamic among the experts. This second step allows us to incorporate the mutual influence between the speaker and the listener.

(LMDE) [9] called mutual-LMDE. LMDE was originally proposed to integrate data from multiple modalities. One of the main advantages of this model is that it can automatically discover the hidden structure among modalities and learn the dynamic between them. We extend the LMDE model to also take into consideration the mutual influence between speaker and listener.

Our mutual-LMDE model is based on a two step process (an overview is shown in Figure 2): in the first step, we learn discriminative experts for speaker gestures and listener backchannels. Speaker expert models are trained using a Conditional Random Field (CRF) [11] on one of the four speech dimensions (prosody, lexicons, syntactic structure and part-of-speech tags). These individual experts make up for the visual context from the speaker. We learn experts for listener backchannels similar to speaker gestures, but using the actual listener backchannel feedback as our labels. In the second step, we merge the speaker experts (visual context) with listener experts by using a latent variable model. This process involves using the outputs of these expert models as an input to a Latent Dynamic Conditional Random Field (LDCRF) [12] that is capable of modeling the mutual influence between listener and speaker gestures.

The task of our LMDE model is to learn a mapping between a sequence of multimodal observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the j^{th} frame of a video sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{$`backchannel`, `no feedback`$\}$. Each frame observation x_j is represented by a feature vector $\in \mathbf{R}^d$, for example, the prosodic features at each sample. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Following Morency et al. [12], we define mutual-LMDE model as follows:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}:\forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h} \mid \mathbf{x}, \theta). \tag{1}$$

where θ are model parameters learned during training and $P(\mathbf{h}| \mathbf{x}, \theta)$ is defined as follows:

$$P(\mathbf{h}| \mathbf{x}, \theta) = \frac{\exp\left(\begin{array}{l}\sum_p \theta_p \cdot \mathbf{T}_p(\mathbf{h}) + \\ \sum_l \theta_l \cdot \mathbf{S}_l(\mathbf{h}, \mathbf{x}) + \sum_s \theta_s \cdot \mathbf{S}_s(\mathbf{h}, \mathbf{x})\end{array}\right)}{\mathcal{Z}(\mathbf{x'}, \theta)}, \tag{2}$$

Different from Ozkan et al. [9] and Morency et al. [12], we learn three sets of θ parameters: (1) θ_p related to the transition between hidden states, (2) θ_l related to *listener* expert outputs, and (3) θ_s related to *speaker* expert outputs. θ_s and θ_l model the relationships between expert outputs and the hidden states h_j. $\mathcal{Z}$ is the partition function. $\mathbf{T}_p(\mathbf{h}, \mathbf{x'})$ is the transition function between the hidden states. $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$ is the listener state function and is defined as follows:

$$\mathbf{S}_l(\mathbf{h}, \mathbf{x}) = \sum_j s_s(h_j, [q_{j_1} q_{j_2} .. q_{j_\alpha} .. q_{j_{|e|}}]) \tag{3}$$

Each q_{j_α} is the marginal probability of expert α at frame j, and equals to $P_\alpha(y_j = a|\mathbf{x}, \lambda_\alpha)$. Each expert conditional distribution is defined by $P_\alpha(\mathbf{y}|\mathbf{x}, \lambda_\alpha)$ using the usual conditional random field formulation:

$$P_\alpha(\mathbf{y}|\ \mathbf{x}, \lambda_\alpha) = \frac{\exp\left(\sum_k \lambda_{\alpha,k} \cdot \mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x})\right)}{\mathcal{Z}_\alpha(\mathbf{x}, \lambda_\alpha)}, \tag{4}$$

where λ_α represent the model parameters of each expert α. $\mathbf{F}_{\alpha,k}$ is either a state function $s_k(y_j, \mathbf{x}, j)$ or a transition function $t_k(y_{j-1}, y_j, \mathbf{x}, j)$. Each expert α contains a different subset of state functions $s_k(y, \mathbf{x}, j)$, defined in Section 4.3.

Speaker state function $\mathbf{S}_s(\mathbf{h}, \mathbf{x})$ is defined similar to $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$. The main difference is that, we use listener backchannels as sequence labels, $\mathbf{y}$, when learning $P_\alpha(\mathbf{y}|\ \mathbf{x}, \lambda_\alpha)$ for listener experts $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$, and use speaker gestures as sequence labels $\mathbf{y}$ for speaker experts $\mathbf{S}_s(\mathbf{h}, \mathbf{x})$.

In our framework, each speaker expert learns a different aspect of speech for speaker gestures. Similarly, the listener experts allows us to obtain discriminative characters of speech for listener backchannel feedbacks. By using a latent variable model to combine these individual experts, our mutual-LMDE model is able to learn both the mutual influence between the speaker and the listener, and the hidden structure among the experts. More details about training and inference of LMDE can be found in Ozkan et al. [9].

4 Experimental Setup

As mentioned in the previous section, we evaluate our mutual-LMDE on the multimodal task of predicting listener nonverbal backchannel. In this section, we first describe the dataset, the gesture and backchannel annotation technique and multimodal speaker features. Then, we explain the baseline models used for comparison in our tests, and the experimental setup.

4.1 Dataset

We are using the RAPPORT dataset from [4], which contains 45 dyadic interactions between a speaker and a listener. Data is drawn from a study of face-to-face narrative discourse ("quasi-monologic" storytelling). In this dataset, participants in groups of two were told they were participating in a study to evaluate a communicative technology. Subjects were randomly assigned the role of speaker and listener. The speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants sat approximately 8 feet apart. All video sequences were manually transcribed and manually annotated to determine the ground truth backchannels. The next section describes our annotation procedure.

4.2 Gesture and Backchannel Annotations

In our experiments, we focus on visual backchannels of a listener: head nods. Similarly, we use speaker head nods as speaker nonverbal behaviors. A head nod gesture starts when the person starts moving his/her head vertically. The head nod gesture ends when the person stops moving or when a new head nod is started. A new head nod starts if the amplitude of the current head cycle is higher than the previous head cycle. Some listeners' responses may be longer than others although they all correspond to one single respond. In our data, annotators found a total of 666 head nods. The duration of these nods varied from 0.16 seconds to 7.73 seconds. Mean and standard deviation of backchannel durations are 1.6 and 1.2 respectively. The minimum number of head nods given by one listener during one interaction is 1, the maximum is 47, mean and standard deviations are 14.8 and 10.9 respectively.

Following Ward and Tsukahara's [14] original work on backchannel prediction, we train our models to predict only the start time of the backchannel start cue (i.e. head nod). Following again Ward and Tsukahara [14], we define the backchannel duration as a window of 1.0 seconds centered around the start time of the backchannel. A backchannel cue will be correctly predicted if at least one prediction of our LMDE model happens during this 1.0 seconds duration. All models tested in this paper use this same testing backchannel duration of 1.0 seconds.

4.3 Multimodal Features and Experts

This section describes the different multimodal audio features used to create our four experts.

Prosody. Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [27, 14, 28]. For example, Ward and Tsukahara [14] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. Listener feedback often follows speaker pauses or filled pauses such as "um" (see [28]). Using openSMILE [29] toolbox, we extract the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara:

- downslopes in pitch continuing for at least 40ms
- regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness)
- drop or rise in energy of speech (i.e., energy edge)
- fast drop or rise in energy of speech (i.e., energy fast edge)
- vowel volume (i.e., vowels are usually spoken softer)
- pause in speech (i.e., no speech)

Lexical. Some studies have suggested an association between lexical features and listener feedback [28]. Using the transcriptions, we included all individual words (i.e., unigrams) spoken by the speaker during the interactions.

Part-of-Speech Tags. In [28], combination of pause duration and a statistical part-of-speech language model is shown to achieve the best performance for placing backchannels. Following this work, we use a CRF part-of-speech (POS) tagger to automatically assign a part of speech label to each word. We also include these part-of-speech tags (e.g. noun, verb, etc.) in our experiments.

Syntactic Structure. Finally, we attempt to capture syntactic information that may provide relevant cues by extracting three types of features from a syntactic dependency structure corresponding to the utterance. The syntactic structure is produced automatically using a data-driven left-to-right shift-reduce dependency parser [30], trained POS on dependency trees extracted from the Switchboard section of the Penn Treebank [31], converted to dependency trees using the Penn2Malt tool[1]. The three syntactic features are:

- Grammatical function for each word (e.g. subject, object, etc.), taken directly from the dependency labels produced by the parser
- Part-of-speech of the syntactic head of each word, taken from the dependency links produced by the parser
- Distance and direction from each word to its syntactic head, computed from the dependency links produced by the parser

Although our current method for extracting these features requires that the entire utterance be available for processing, this provides us with a first step towards integrating information about syntactic structure in multimodal prediction models. Many of these features could in principle be computed incrementally with only a slight degradation in accuracy, with the exception of features that require dependency links where a word's syntactic head is to the right of the word itself. We leave an investigation that examines only syntactic features that can be produced incrementally in real time as future work.

4.4 Baseline Models

Individual Experts. Our first baseline model consists of a set of CRF chain models, each trained with different set of multimodal features (as described in the previous section). In other words, only visual, prosodic, lexical or syntactic features are used to train a single CRF expert. (See Figure 3a).

Multimodal Classifiers. Our second baseline consists of two models: CRF and LDCRF [12]. To train these models, we concatenate all multimodal features (lexical, syntactic and prosodic) in one input vector. Graphical representation of these baseline models are given in Figure 3-(a) and Figure 3-(b).

[1] `http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html`

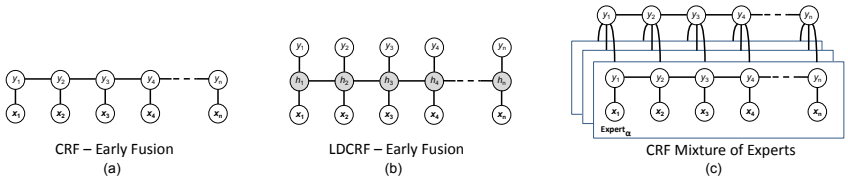

Fig. 3. Baseline Models: **a)** Conditional Random Fields (CRF), **b)** Latent Dynamic Conditional Random Fields(LDCRF), **c)** CRF Mixture of Experts (no latent variable)

LMDE. To show the importance of visual context from the speaker, we train an LMDE model without using any of the speaker experts. In other words, our baseline LMDE model is trained to directly predict listener backchannels from the speaker audio features.

Pause-random Classifier. Random backchannel generator randomly generates backchannels whenever some pre-defined conditions in the speech is purveyed. These conditions include pauses that come after at least 700 milliseconds of speech and absence of backchannel feedback within the preceeding 800 milliseconds. This random classifier has also been used by Ward and Tsukahara [14] for comparison.

CRF Mixture of Experts. To show the importance of latent variable in our context-based prediction model, we trained a CRF-based mixture of discriminative experts. A graphical representation of a CRF Mixture of experts is given in Figure 3. This model is similar to the Logarithmic Opinion Pool (LOP) CRF suggested by Smith et al. [13], in the sense that they both factor the CRF distribution into a weighted product of individual expert CRF distributions. The main difference between LOP and CRF Mixture of Experts model is in the definition of optimization functions. Training of CRF Mixture of Experts is performed in two steps: Expert models are learned in the first step, and the second level CRF model parameters are learned in the second step.

LMDE with Speaker Nods. Our final set of baseline models include an LMDE model that directly uses the visual context from the speaker (speaker nods). In this baseline model, we first train only the listener expert models as in the first step of our proposed approach. Then, in the second step, we use the annotated (actual) speaker gestures together with the listener experts as input to the latent variable model. So, the main difference of this baseline model with our approach is that our approach first anticipates the speaker nonverbal behaviors through CRF experts instead of directly using them.

4.5 Methodology

We performed held-out testing by randomly selecting a subset of 11 interactions (out of 45) for the test set. The training set contains the remaining 34 dyadic

interactions. All models in this paper were evaluated with the same training and test sets. Validation of all model parameters (regularization term and number of hidden states) was performed using a 3-fold cross-validation strategy on the training set. The regularization term was validated with values $10^k, k = -1..3$. Two different number of hidden states were tested for the LDCRF models: 3, and 4 (note that LDCRF with 1 hidden state is equivalent to Mixture of CRF Experts model).

The performance is measured by using the conventional metrics: precision, recall, and F-measure. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model. We use the same weight for both precision and recall, so-called F_1, which is the weighted harmonic mean of precision and recall. F_1 scores for each sequence is calculated first, then the final F_1 result is computed by averaging these sequence scores.

Before reviewing the prediction results, is it important to remember that backchannel feedback is an optional phenomena, where the actual listener may or may not decide on giving feedback [14]. Therefore, results from prediction tasks are expected to have lower accuracies as opposed to recognition tasks where labels are directly observed (e.g., part-of-speech tagging).

During testing, we find all the "peaks" (i.e., local maxima) from marginal probabilities $P(y_j = a|\ \mathbf{x}, \theta)$. For the f1-score, the prediction model needs to decide on a specific threshold (i.e., amount of backchannel) for the marginal probabilities for all users. The value of this threshold is automatically set during validation. Since we are predicting the start time of a backchannel, an actual listener backchannel is correctly predicted if at least one model prediction happen within the 1 second interval window around the start time of the listener backchannel.

The training of all CRFs and LDCRFs were done using the hCRF library[2]. The LMDE model was implemented in Matlab based on the hCRF library. The input observations were computed at 30 frames per second. Given the continuous labeling nature of our model, prediction outputs were also computed at 30Hz.

5 Results

In this section we present the results of our empirical evaluation. We designed our experiments to test different characteristics of our mutual-LMDE approach: (1) integration of multiple sources of information, and (2) mutual influence.

Performances of individual CRF experts for predicting listener backchannels and speaker gestures are presented in Table 1. Our approach combines all these experts to model the mutual influence between the speaker and the listener. This integration of multiple resources improve the prediction accuracy for listener backchannels. Therefore, we get an f-1 score of 0.32 with our mutual-LMDE model.

[2] `http://sourceforge.net/projects/hrcf/`

Table 1. Test performances of the individual expert models for listener backchannel and speaker gesture (head nod) predictions

	Listener			**Speaker**		
Expert	f1	Precision	Recall	f1	Precision	Recall
Prosodic	0.1913	0.1060	0.9803	0.2789	0.1669	0.8478
Lexical	0.2073	0.1377	0.4198	0.2959	0.2068	0.5203
POS	0.2346	0.1446	0.6220	0.3274	0.2182	0.6556
Syntactic	0.2045	0.1287	0.4956	0.3175	0.2330	0.4983
mutual-LMDE	**0.3212**	0.2633	0.4117	**0.3313**	0.2456	0.5087

Table 2. Comparison of different models with our approach

Model	f1	Precision	Recall
Early CRF	0.2173	0.1423	0.4591
Early LDCRF	0.2115	0.1231	0.7495
LMDE	0.2764	0.2055	0.4219
Pause-Random	0.1456	0.1322	0.2031
CRF Mixture	0.1963	0.1718	0.2288
LMDE+Speaker Nods	0.2614	0.2071	0.3541
mutual-LMDE	**0.3212**	0.2633	0.4117

In our second set of experiments, we evaluate the importance of modeling mutual influence. Table 2 summarizes our results. The prediction models in the top three rows of the table do not take into account the mutual influence between the speaker and the listener. These models are trained on the speaker audio features to directly infer the listener backchannels. Among these models, LMDE gives the best f-1 score, which proves the importance of late fusion of multiple sources of information (different speech channels). However, our mutual-LMDE model outperform all these three models, which indicates the importance of using mutual influence between the interlocutors.

The models listed in the last three rows of Table 2, model the mutual influence. CRF Mixture model does not perform as good as other LMDE models. The main reason for this decrease in performance is that the LMDE model uses a latent variable to capture the dynamic among different sources of information, whereas the CRF Mixture approach directly models these information. Although the last LMDE approach use the speaker nonverbal behavior information directly in the second step of LMDE, it does not perform as good as our mutual-LMDE model, in which we first infer these speaker behaviors instead of directly using them. We hypothese that, by inferring the speaker backchannels, we are able to model a better average speaker feedback behavior and remove the variations in the actual speaker backchannels.

Our framework addresses the problem of listener backchannel prediction by modeling the mutual influence. A related issue is modeling the recursive influence between the listener and the speaker. For instance, backchannels of a

listener might trigger more visual gestures from the speaker. Although we do not explicitly model this recursive influence in our current study, the proposed framework can be extended to address this issue as well. For instance, we can use the listener observations (features) in the learning process for speaker experts to model how listener behaviors affect speaker behaviors. The study of these recursive models is part of our future work.

6 Conclusions

In this paper, we proposed a context-based approach for predicting the backchannels of a listener in a dyadic conversation. To model the mutual influence between the speaker and the listener, we used a variant of Latent Mixture of Discriminative Experts model. Our mutual-LMDE approach consists of two steps: we first learn expert models to predict speaker gestures (head nods), and the listener backchannel feedbacks. Then, we use visual context (predicted speaker gestures) from the speaker to improve the final listener backchannels.

We evaluated our approach on 45 dyadic interactions from the RAPPORT dataset. Our experiments have shown improvement over all previous approaches. The results suggest two main conclusion: (1) By modeling the mutual influence between the participants of a dyadic interaction, we can better model the backchannel feedbacks of the listener. (2) In case of no available visual speaker information, predicted speaker visual context helps us to learn an average speaker behavior that is more effectual and less noisy than actual speaker behaviors.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Ross, M.D., Menzler, S., Zimmermann, E.: Rapid facial mimicry in orangutan play. Biol. Lett. 4, 27–30 (2008)
2. Hatfield, E., Cacioppo, J., Rapson, R.: Emotional contagion. In: Clark, M.S. (ed.) Review of Personality and Social Psychology: Emotion and Social Behavior, pp. 151–171 (1992)
3. Riek, L.D., Paul, P.C., Robinson, P.: When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. Journal on Multimodal User Interfaces 3, 99–108 (2010)
4. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
5. Drolet, A.L., Morris, M.W.: Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Journal of Experimental Social Psychology 36(1), 26–50 (2000)

6. Tsui, P., Schultz, G.: Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients. American Journal of Orthopsychiatry 55, 561–569 (1985)
7. Fuchs, D.: Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education 7, 90–104 (1987)
8. Burns, M.: Rapport and relationships: The basis of child care. Journal of Child Care 2, 47–57 (1984)
9. Ozkan, D., Morency, L.P.: Latent mixture of discriminative experts. IEEE Transactions on Multimedia 15(2), 326–338 (2013)
10. Morency, L.P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Conference on Intelligent Virutal Agents, IVA (2008)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labelling sequence data. In: International Conference on Machine Learning, ICML (2001)
12. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEE Conference on Computer Vision and Pattern Recognition, CVPR (2007)
13. Smith, A., Cohn, T., Osborne, M.: Logarithmic opinion pools for conditional random fields. In: Association for Computational Linguistics (ACL), pp. 18–25 (2005)
14. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in english and japanese. Journal of Pragmatics 23, 1177–1207 (2000)
15. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. In: ACM International Conferance on Multimodal Interfaces, pp. 239–248 (2006)
16. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)
17. Sebea, N., Cohenb, I., Netherl, T.: Multimodal approaches for emotion recognition: A survey (2005)
18. Maatman, R.M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
19. Nakano, Y., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: Association for Computational Linguistics, ACL (2003)
20. Nakano, Y., Murata, K., Enomoto, M., Arimoto, Y., Asa, Y., Sagawa, H.: Predicting evidence of understanding by monitoring user's task manipulation in multimodal conversations. In: Association for Computational Linguistics (ACL), pp. 121–124 (2007)
21. Ward, N.: Non-lexical conversational sounds in American English (2003)
22. Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., Kobayashi, T.: A conversation robot using head gesture recognition as para-linguistic information. In: IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 159–164 (2004)
23. Kang, S.H., Gratch, J., Wang, N., Watt, J.: Does the contingency of agents' nonverbal feedback affect users' social anxiety? In: International Conference on Autonomous Agents and Multiagent Systems, AAMAS (2008)
24. Semaine the sensitive agent project
25. Gravano, A.: Turn-taking and affirmative cue words in taskoriented dialogue. Technical report (2009)

26. Neiberg, D.: Modelling Paralinguistic Conversational Interaction: Towards social awareness in spoken human-machine dialogue. PhD thesis, KTH, Speech Communication and Technology, QC 20120914 (2012)
27. Nishimura, R., Kitaoka, N., Nakagawa, S.: A spoken dialog system for chat-like conversations considering response timing. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 599–606. Springer, Heidelberg (2007)
28. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: European Chapter of the Association for Computational Linguistics (EACL), pp. 51–58 (2003)
29. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Affective Computing and Intelligent Interaction (ACII), pp. 576–581 (2009)
30. Sagae, K., Tsujii, J.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Association for Computational Linguistics (ACL), pp. 1044–1050 (2007)
31. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The penn treebank: annotating predicate argument structure. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 114–119 (1994)