Speaker-Adaptive Multimodal Prediction Model for Listener Responses

Iwan de Kok
Human Media Interaction
University of Twente
Enschede, The Netherlands
i.a.dekok@utwente.nl

Dirk Heylen
Human Media Interaction
University of Twente
Enschede, The Netherlands
heylen@utwente.nl

Louis-Phillippe Morency
USC Institute for Creative Technologies
Los Angeles, United States
morency@ict.usc.edu

ABSTRACT

The goal of this paper is to acknowledge and model the variability in speaking styles in dyadic interactions and build a predictive algorithm for listener responses that is able to adapt to these different styles. The end result of this research will be a virtual human able to automatically respond to a human speaker with proper listener responses (e.g., head nods). Our novel speaker-adaptive prediction model is created from a corpus of dyadic interactions where speaker variability is analyzed to identify a subset of prototypical speaker styles. During a live interaction our prediction model automatically identifies the closest prototypical speaker style and predicts listener responses based on this communicative style. Central to our approach is the idea of "speaker profile" which uniquely identify each speaker and enables the matching between prototypical speakers and new speakers. The paper shows the merits of our speaker-adaptive listener response prediction model by showing improvement over a state-of-the-art approach which does not adapt to the speaker. Besides the merits of speaker-adaptation, our experiments highlights the importance of using multimodal features when comparing speakers to select the closest prototypical speaker style.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents*

General Terms

Algorithms, Human Factors, Theory

Keywords

Listener Responses, Machine Learning, Social Behavior, Multimodal

ICMI'13, December 9–13, 2013, Sydney, Australia

ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522866.

1. INTRODUCTION

During face-to-face conversation people naturally coordinate through their verbal and nonverbal behaviors. This multimodal coordination is utilized to regulate turn-taking, emphasize important parts of the interaction, establish rapport with the interlocutors, among other things. It is a constant back and forth where actions are chosen depending on the behaviors of the other interlocutor(s). The coordination between interlocutors shows in their speech through changing voice levels, utterance frequency and pauses [16], as well as visual behaviors such as postures, facial expressions and other gestures [6].

This collaborative coordination occurs both while speaking and listening [3]. While listening interlocutors give so called listener responses (e.g., head nods or short vocalizations like "uh-huh" and "okay"). These listener responses are optional, but are placed at specific places in the discourse. Oftentimes the speaker cues these places and expects a listener to respond [15]. The absence of the expected listening behavior at such places can result in restarts (and often rephrases) from the speaker [13]. This affects the fluency of the conversation, which in turn affects speaker clarity and ultimately speaker comprehension [22, 3]. It has also been proven to hurt the rapport between interlocutors [14].

Our long-term goal is to create an embodied conversational agent that is capable of having a natural conversation with a human. Appropriate listening behavior is a key component in such an agent. To be able to generate listening behavior, the agent needs to be able to identify the moments where a listener response is appropriate based on observations of the verbal and nonverbal behavior of the speaker. In this paper we call a model performing this task a listener responses prediction model.

Since the first listener response prediction model was proposed in 1989 [31] many have followed (see Section 2). A key observation not explicitly modeled in prior approaches is the variability in speaker styles and personalities. Prior work in conversation analysis focussed on finding similarities in speaker behavior in relation to listener responses (see [3, 15, 30]). For instance, it is known that looking towards the listener at the end of a sentence is a good cue for predicting listener responses [3, 25]. However, not every person is as comfortable with looking other people in the eye during conversations as others and they will do this less often. When a prediction model used by a virtual agent is heavily dependent on this cue, this prediction model will probably not perform as well for this speaker.

In this paper we introduce a speaker-adaptive listener response prediction model which takes into consideration the variability of speaking styles. Our speaker-adaptive model is created from of a collection of dyadic speaker-listener interactions. Our prediction model identifies a subset of prototypical speakers and creates prediction models for each of them. When encountering a new speaker our model analyzes the characteristics of the speaker and selects the prediction model that reflects similarities with our prototypical speakers.

A key challenge in our approach is to find a representation of the speaker behaviors that highlights the differences between prototypical styles while acknowledging their similarities. We name this representation a "speaker profile" and it will be a central component used to match new speakers with their closest prototype.

An extensive set of experiments are presented on the MultiLis corpus [8] and a comparison is made between our approach and previously published models on the same dataset. Besides the merits of speaker-adaptation, our experiments highlight the importance of using multimodal speaker profile when comparing speakers to select the appropriate model matching the speaking style of the current interlocutor.

The paper continues in Section 2 with a presentation of previous work on listener response prediction models and user-adaptive modeling. Section 3 describes our approach to the speaker-adaptive listener response prediction model in more detail. The experiment to evaluate the proposed model is presented in Section 4. The results of this experiment are presented and discussed in Section 5. The paper concludes and presents future directions for our work in Section 6.

2. RELATED WORK

Since the first handcrafted listener response prediction model was proposed in 1989 by Watanabe and Yuuki [31] many have followed. In general, these models are difficult to compare in terms of performance as they are created and tested on different corpora and present varying evaluation metrics [9].

The first machine learning approach was proposed by Okato et al. [26]. They learned a Hidden Markov Model to detect prosodic patterns that can predict listener responses. Ward and Tsukahara [30] proposed 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.

Maatman et al. [24] presented the first multimodal approach. In their approach they combined Ward and Tsukhara's prosodic algorithm 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 [19] and speech fluency [14]. The first multimodal machine learning approach was presented by Morency et al. [25]. They used Conditional Random Fields to learn a listener response prediction model and showed statistical improvement when compared to the handcrafted approach of Ward and Tsukahara [30]. Given its wide applicability on other datasets, this approach was used as baseline for this paper.

Since then, the main focus has shifted to increase performance by collecting listener responses from more listeners to get a wider coverage of response opportunities. De Kok et al. [10] recorded multiple listeners in interaction with the same speaker. Huang et al. [17] collected listener responses through parasocial sampling, where listeners watch prerecorded videos of a speaker and give listener responses through the keyboard as if they were listening. These additional listener responses proved to improve performance of the prediction models. Both researchers learned models from the consensus between the listeners, thus ignoring individuality of the interlocutors.

Ozkan and Morency [27] used parasocial sampling to collect listener responses from nine 'parasocial' listeners on 43 interactions. Subsequently nine expert prediction models were learned using Conditional Random Fields, one for each listener. The output of these expert models served as input for a Latent Dynamic Condition Random Field that combined the knowledge captured in the experts.

A closely related field to our approach is domain adaptation in the natural language processing community. In this field domain adaptation is achieved by adjusting a model learned on a specific dataset (domain) to match the data distribution of the new domain. Recognition of which features are important can be achieved online. This online learning/reweighting technique has been succesfully applied to adjust to speakers in the dialogue act recognition task [28].

To the best of our knowledge the listener response prediction model proposed in this paper is the first model that explicitly adapts to the variability of speakering styles.

3. SPEAKER-ADAPTIVE PREDICTION MODEL OF LISTENER RESPONSES

In this section we introduce our speaker-adaptive prediction model of listener responses. The section starts with a general description of our prediction model. Later we will explain the main novelty of our model, speaker adaptation through selection of the listener response prediction model based on the speaker profiles, in more detail. Finally, we describe our method for selecting models for inclusion in the model collection.

3.1 Overview

An important steop of our speaker-adaptive prediction model is the model collection. The model collection consists of prototypical listener response prediction models that represent the variability in speaking styles found in the corpus. How the prediction models in the model collection represents the variability in speaking style is described in more detail in section 3.3.

The models in the model collection are learned on individual speaker-listener pairs from a corpus of dyadic interactions during the offline learning phase. Each individual model learns the mapping between the features that are extracted from the audio and video signal of the speaker and the ground truth labels that represent the times at which the listener has given a listener response in the corpus. With each model the speaker profiles are included in the model collection. These speaker profiles describe the speaking style the model represents. More details on the speaker profile follows in Section 3.2.

The online prediction cycle is depicted in Figure 1. When a new speaker is encountered, the speaker profile describing the behavior of this speaker are calculated. This speaker

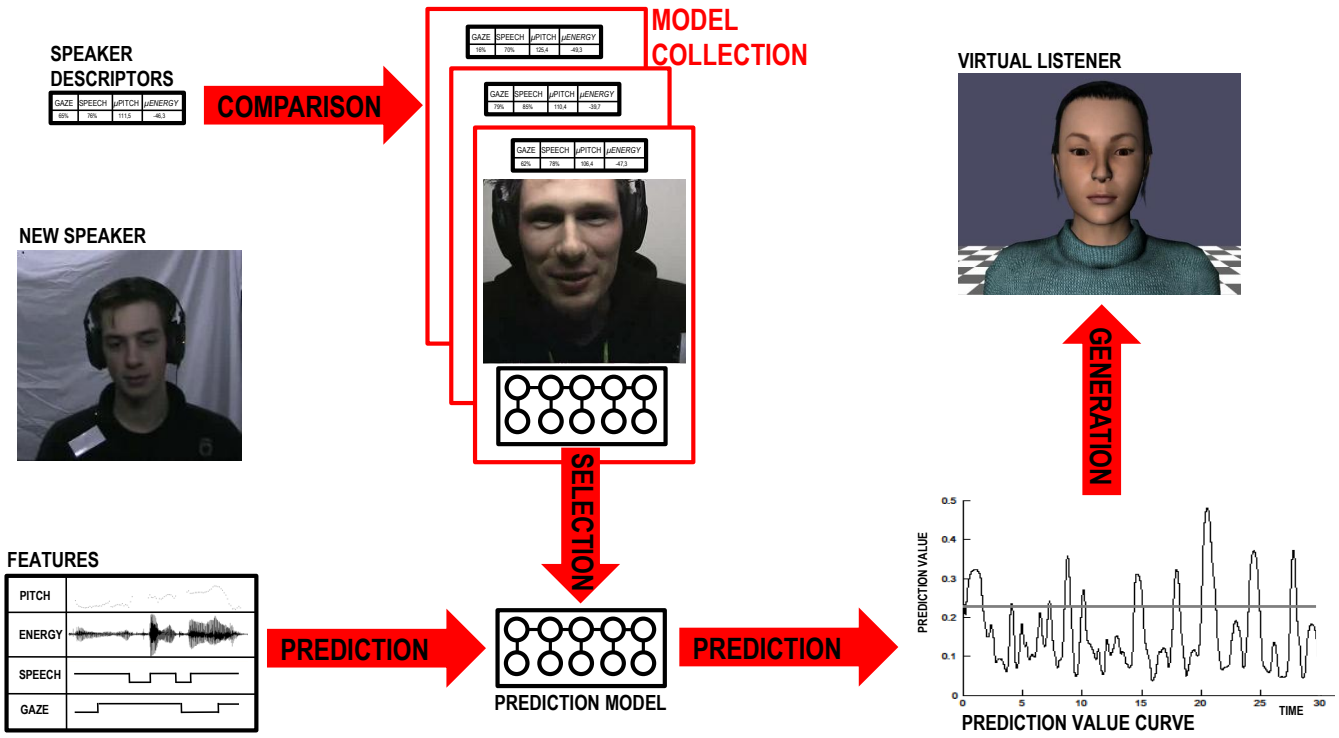

Figure 1: The figure illustrates the online prediction cycle for our speaker-adaptive prediction model. The model collection includes prediction models learned on individual speaker-listener pairs and a speaker profile describing the speaking style. When encountering a new speaker the speaking style of this speaker is compared to the speaking styles of all speaker in the model collection through the speaker profiles. The model associated with closests matching speaker profile is selected to predict the listener responses for the virtual listener.

profile is compared to all speaker profiles in the model collection, depicted in the center of the figure. The model associated with the nearest-neighbor match measured on the speaker profiles is selected. Thus, the selected model is the model that is learned on an interaction that is the most similar to the interaction the model is engaged in currently.

The model is applied to the extracted features of the new speaker which results in a prediction value curve with the probability of a listener response at each time frame. Using this prediction value curve the listening behavior of the virtual human is generated.

3.2 Speaker Profiles

One of the novel challenges the speaker-adaptive listener response prediction model introduces is the challenge to find a similar speaker based on multimodal features. A closely related challenge is speaker diarization, where a group of speakers need to be discriminized into individual speakers [29]. However, our challenge is not finding the exact same speaker among others, but finding a speaker with a similar speaker style that cues the moments where he/she expects a listener response in a similar way. Little is known about how speakers differ in cueing listener response opportunities. Similar to the development of listener prediction models, conversation analysis literature has also focussed on findings by pooling all speaker and listener pairs from the corpus together and finding similarities.

Features that are often found in conversation analysis literature to be associated with listener response opportunities include the pitch [21, 30, 15] and energy [21, 15] of the speech signal, pauses in speech [11, 5] and the eye gaze of the speaker [20, 2, 4]. Therefore, it is to be expected that differences lies in these same features. Thus, our focus for the speaker profiles was directed towards these features.

Each speaker profile consists of several speaker descriptors. A speaker descriptor summarizes the behavior of the speaker during the whole interaction for a certain feature in a single value. For features that are a continuous signal (e.g. pitch and energy) the speaker descriptors are the mean and standard deviation of the signal. For binary features (e.g. speech segments and eye gaze) the speaker descriptors are percentage of true values and number of segments per minute.

To select a prediction model the speaker profile are compared to all speaker profiles in the model collection. There are many ways to compare two vectors and find the closest match. For our model nearest-neighbor measured by Euclidean distance was chosen. The results presented in Section 5 will show that speaker adaptation on these basic speaker profiles and straight forward nearest neighbor selection will improve prediction performances compared to the a state-of-the-art CRF model without speaker adaptation.

3.3 Model Collection Composition

As stated before the speaker profiles are used to select a model from the model collection. Each prediction model in the model collection represents a different speaking style. However, not every model trained on a speaker is suited for inclusion into the model collection. The composition of the model collection is a balance act between 1) the quality of the individual prediction model and 2) the contribution to the representation of variability in speaking style. In other words, the goal of the model collection is to have a representative model for as many different speakers as possible.

This does not necessarily mean that adding as many individual models as possible to the model collection improves the performance of the speaker-adaptive prediction model. If the model collection already includes a good prediction model for a similar speaker, it is better to use that model as a representative for the speaker, than an inferior model. Therefore, models included in the model collection are selected based on their individual performance, while controlling for representation of the variability in speaking style.

4. EXPERIMENT

In this section the experiment that has been conducted will be reported. The goals of our experiments are to (1) compare our speaker-adaptive approach with priori state-of-the-art approaches, and (2) study the effect of the each of the different modalities in our speaker profiles.

The section will start with a description of the MultiLis corpus that is used for learning and evaluating the models. This will be followed by the details of learning the models. After this, the details of the model selection in the user-adaptive learning approach will be described. Finally, the details of the evaluation will be presented.

4.1 Corpus

The publicly available MultiLis corpus [8] was used for the learning and evaluation of our listener response prediction models. The corpus consists of 32 Dutch-spoken mediated human-human interactions between pairs of subjects. In the first interaction, one subject assumed the role of speaker and one subject was assigned the role of listener. In a second interaction, the roles were switched. In total, 32 subjects (29 male, 3 female, mean age 25) participated in 32 recordings, with a total duration of 131 minutes for an average of little over 4 minutes per interaction.

The speakers were instructed to either summarize a short video or to provide the instructions of a recipe they had just studied. Listeners had to remember as many details as possible. Subjects interacted through a remote videoconferencing system. The camera was placed behind an interrogation mirror on which the other subject was projected. This allowed subjects to look directly at the camera and this created the feeling of eye contact. In addition, this setting allowed us to analyze gaze.

The onsets of the 886 listener responses found in the corpus are manually annotated. The listener responses consist of 90% head nods and the remaining 10% are short vocalizations such as “uh-huh” and “okay”.

4.2 Model Learning

The machine learning models trained in our experiments are Conditional Random Fields (CRF) [23] and are trained using the hCRF library [1]. CRF is a probabilistic discriminative model for sequential data labeling. A CRF learns a mapping between a sequence of observations, in this case the learning features describing the behavior of the speaker, and a sequence of ground truth labels, in this case the onsets of listener responses from the MultiLis corpus as positive samples and the same amount of randomly selected moments where no listener response occured as negative samples. The learned model returns a prediction value curve with a value at each frame indicating the probability of a listener response. After smoothing the prediction value curve can be used to predict listener responses by detecting peaks in the curve. By comparing the heights of these peaks to a threshold the most probable moments are selected as predicted response opportunities.

In this experiment two models are compared, the baseline model and a model using the technique explained in Section 3. For this comparison the following models were learned:

- **State-of-the-art CRF Model** - Thirty-two state-of-the-art CRF models are learned. Each of these models is learned using 31 interactions from the MultiLis corpus as learning data and the remaining interaction as test data.
- **Individual Models** - Thirty-two individual models are learned. Each of these models is learned using one interaction from the MultiLis corpus as learning data and the remaining 31 interactions as test data. A subset of these individual models are selected for the model collection of our speaker-adaptive multimodal prediction model (see Section 4.4).

The comparison was made using a 32-fold or leave-one-out cross validation at the interaction level. For each validation fold one interaction was left out of the training set for the baseline model. For the proposed speaker-adaptive model, the individual model that was learned on this interaction was unavailable to be included in the model collection.

All models are learned on the learning features. These features describe the behavior of the speaker on a frame by frame basis at a frequency of 25 Hz. There are six features, of which four are acoustic features, one is a turn-taking feature and one a visual feature. These features are:

- **Pitch** - The raw pitch values were extracted using the algorithm of Drugman and Alwan [12] at a sampling rate of 100 Hz. Gaps in detected pitch smaller than 80 ms (8 frames) are linearly interpolated, following [30]. Then all pitch values are converted to their z-score equivalent. Afterwards the feature is downsampled to 25Hz.
- **Pitch Slope** - As a measurement of the change of the pitch the slope of the pitch value feature is calculated by taking the first derivative of the pitch signal.
- **Energy** - The energy of each speech frame is calculated on 32 ms Hanning windows with a shift of 10 ms and expressed in dB.
- **Energy Slope** - As a measurement of the change in speech intensity the slope of the energy value feature is calculated by taking the first derivative of the energy signal.

- **Speech Segment** - The speech segment feature captures whether the speaker is speaking at the moment or not. It is represented as a binary feature. The feature is extracted using the segmentation from the Dutch automatic speech recognizer SHoUT [18]. The minimum pause between speech segments is 100ms (4 frames).

- **Gaze** - The gaze feature is represented as a binary feature that is true when the speaker looks directly at the listener. The feature is extracted from the annotations provided in the MultiLis corpus.

4.3 Speaker Profiles

Our speaker-adaptive model has a model collection. This model collection includes the models that are learned from single speaker-listener pairs. With each model a description of the behavior of the speaker of the pair the model is inferred from is associated. The behavior is captured in a speaker profile containing 10 speaker descriptors. The speaker descriptors summarize the behavior of the speaker over the course of the interaction. Our speaker descriptors include six acoustic features, two turn-taking features and two gaze features. These are:

- **Mean Pitch** - The mean of all Pitch values of the interaction. The pitch values are the values from before converting to the z-score equivalent (otherwise the mean would always be 0).

- **Standard Deviation of Pitch** - The standard deviation of all Pitch values of the interaction. Again using the raw pitch values before converting to the z-score equivalent (otherwise the mean would always be 1).

- **Mean Energy** - Mean of all Energy values of the interaction expressed in dB.

- **Standard Deviation of Energy** - Standard deviation of all Energy values of the interaction.

- **Mean Energy Slope** - Mean of all Energy Slope values of the interaction.

- **Standard Deviation of Energy Slope** - Standard deviation of all Energy Slope values of the interaction.

- **Percentage of Speech** - The percentage of time the speaker is speaking.

- **Speech Segments per Minute** - The number of speech segments per minute.

- **Percentage of Gaze** - The percentage of time the speaker is looking at the listener.

- **Gaze Shifts per Minute** - The number of gaze shifts per minute.

When encountered with a new speaker our speaker-adaptive model calculates the speaker profile for the new speaker and compares it to the speaker profiles found in the model collection. It selects the model whose speaker profile is the nearest neighbor match as measured by the euclidean distance.

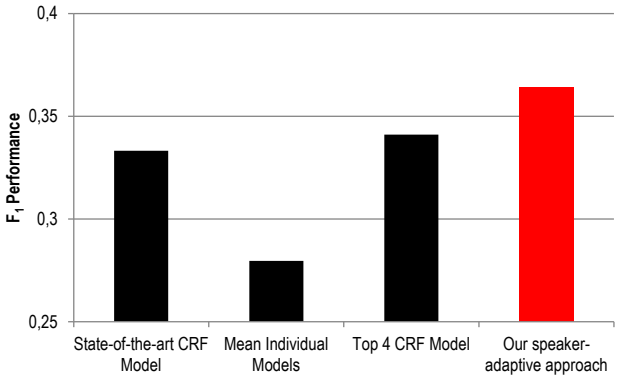

Figure 2: The figure illustrates the performance of the models included in the experiment. The model proposed in this paper is presented in red and the models it is compared to in black. The figure illustrates that our speaker-adaptive model performs best with a performance of 0,364. The difference between our speaker-adaptive model and the state-of-the-art CRF model is significant ($t(31) = 3,25, p = 0,001$).

4.4 Model Collection Composition

As previously stated the composition of the model collection is a balance act between 1) the quality of the individual prediction model and 2) the contribution to the representation of variability in speaking style. The composition of the model collection is based on the performance of the individual models. To find the optimal model collection the number of models included in the model collection was varied from N=1 to N=31. With each collection size the top N models were selected based on individual performance.

Afterwards the representation of variability was controlled for by placing each speaker in the 2D space drawn up by the first two principal components of the speaker profiles.

4.5 Evaluation

The models are evaluated by comparing the predictions made by the model to the listener responses found in the MultiLis corpus.

Predictions are made by selecting the peaks from the prediction value curve that exceed a certain threshold. Usually, e.g. [25, 10, 27], this threshold is determined on the learning set during a validation phase. However, this method for determining the threshold is unreliable. For some models the threshold is set too low, resulting in too many predictions, while for others the threshold is set too high, resulting in no predictions. Especially, for the individual models our learning set is very limited which makes the validated threshold unreliable. To not be dependent on this, the threshold is optimized such that it gives us the optimal performance on each interaction during testing. This is done for all models.

Performance is measured using the F_1 measure. This measure is the weighted harmonic mean of precision and recall. A prediction is considered a true positive if it is made within 500 ms from the onset of a listener response found in the MultiLis corpus. The performances of the models in the same conditions are averaged.

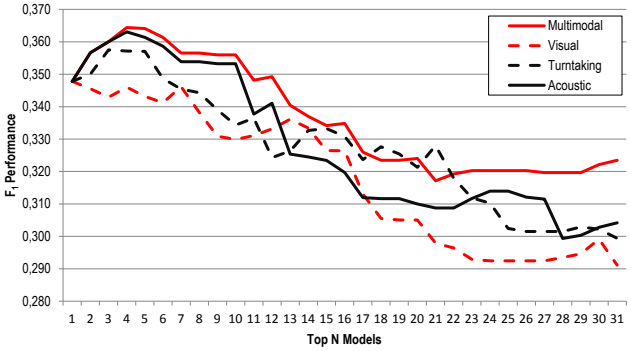

Figure 3: The figure illustrates two points. First, the maximum performance is achieved by including the top 4 performing models in the model collection with a performance of 0,364. Second, the figure illustrates the importance of the multimodality in comparing speakers. The multimodal nearest neighbor selection (solid red line) almost always outperforms the unimodal nearest neighbor selection.

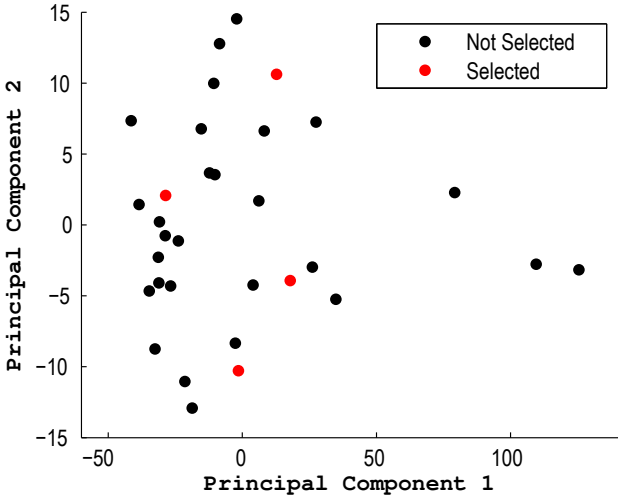

Figure 4: The figure places each speaker in the 2D space created by the first two principal components of the speaker profiles. The figure illustrates that the four models that are selected for the model collection (red) are good representatives of the diversity found in the speakers, since they are spread out over the 2D space.

5. RESULTS

In this section the results of the experiments are presented. The section will be started with presentation of the increase in performance our speaker-adaptive multimodal listener response prediction model achieves over the state-of-the-art CRF model in Section 5.1. This will be followed by an analysis of the importance of the model collection composition 5.2. Finally, the importance of multimodality of the speaker profiles will be analyzed in Section 5.3.

5.1 Speaker-Adaptation

The performances of the models in question are presented in Figure 2. In this figure the performances presented in red are for the models proposed in this paper, while the black models are the performances of the models these proposed models are compared to.

The performance of our speaker-adaptive listener responses prediction model is a F_1 score of 0,364 (fourth bar in Figure 2). This is better that average performance of the state-of-the-art CRF model, which has a performance of a F_1 score of 0,333 (first bar in Figure 2). This difference is significant, $t(31) = 3,25, p = 0,001$.

Our speaker-adaptive model has a model collection of individual models. The average performance of these individual models is a F_1 score of 0,280 (second bar in Figure 2). The best individual model performs at a F_1 score of 0,348. The model collection of our best speaker-adaptive model includes four the top 4 individual models (see for more details on the selection process Section 5.2). The average performance of these four top 4 individual models is 0,342. A state-of-the art CRF model that is learned using the top 4 four interactions that are used as learning data for these individual models performs at a F_1 score of 0,341 (third bar in Figure 2). The fact that our speaker-adaptive model performs better than this model proves that the speaker adaptation accounts for most of the performance boost and not only the characteristics of the learning data.

5.2 Model Collection Composition

As previously stated the composition of the model collection is a balance act between 1) the quality of the individual prediction model and 2) the contribution to the representation of variability in speaking style. In this section the importance of the composition of the model is analyzed in more detail.

To find the optimal model collection the number of models included in the model collection was varied from N=1 to N=31. The composition of the model collection was determined by selecting the top N individual models based on the mean performance as measured by the F_1 score. In Figure 3 the results of varying the number of models in the model collection is presented by the solid red line. The other lines are discussed in Section 5.3.

The figure illustrates the maximum performance is achieved when the top 4 models are included in the model collection. At this number of the models the performance peaks at 0,364 (right bar in Figure 2). The speaker-adaptive model that includes all individual models in the model collection gives a performance of a F_1 score 0,323. This is worse than both the state-of-the-art CRF model and the best individual model. The inclusion of some of the individual models hurts our performance. These results highlight the importance of the composition of the model collection.

Limiting the model collection to only the top 4 models might have caused the model collection to be less representative of the variability in speaking styles than desired. The idea behind the model collection is to have a close match for any new speaker the model may encounter. However, since the models included in the model collection are purely selected on their performance, the top 4 models might actually be close neighbors to each other in the speaker profile space.

To analyze this a principal component analysis was made on the speaker profiles. The first two principal components, which account for 96,2% of the variability, are selected and each speaker is placed in the 2D space that these components create. The results of this analysis is presented in Figure 4.

In this figure the four speakers that are selected for the model collection are plotted in red and the remaining 28 speakers in black. The figure illustrates that the four speakers are well spread out over the 2D space. Thus, the models are a good representative of the variability in speaking styles found in the MultiLis corpus.

5.3 Multimodal Speaker Profile

Finally, the importance of multimodality of our speaker profiles was analyzed. A comparison was made between speaker profiles with multimodal speaker descriptors and unimodal speaker descriptors of the three modalities (acoustic, visual and turntaking). The comparisons were made on the speaker-adaptive models with varying model collection compositions developed for the previous analysis in Section 5.2. The results are presented in Figure 3.

Our speaker-adaptive listener response prediction model with multimodal speaker profiles is represented by the solid red line. For almost all model collection compositions the multimodal speaker descriptors outperform the unimodal speaker profiles. For the best models the acoustic speaker profiles (solid back line) contribute the most to the performance. However, it are actually the turn-taking speaker profiles that outperform the multimodal speaker profiles for some model collection compositions (N=18,19 and 21).

6. CONCLUSION AND FUTURE WORK

In this paper a speaker-adaptive model for predicting listener responses is presented. This speaker-adaptive model consists of a collection of listener response prediction models that are trained on single interlocutor pairs. The composition of this model collection represents the variability in speaker styles found in the corpus as measured by the speaker profiles. When encountering a new speaker the model compares the speaker profile of this speaker to all the speaker profiles in the model collection. The model that is learned on the closest matching speaker is used to predict listener response opportunities for the new speaker.

As reported in Section 5 the performance of this model was compared to a state-of-the-art CRF model. Our approach proved to outperform the state-of-the-art approach as measured by the F_1 measure (0,333 for the baseline model versus 0,364 for our speaker-adaptive model). The performance is also comparable to the F_1 scores achieved when comparing humans interacting with the same speaker to each other (between 0.18 and 0.52 [7]). Experiments showed that the speaker-adaptation, the composition of the model collection and the multimodality of the speaker profiles are all important factors contributing to the performance of our approach.

The presented model opens exciting new avenues for future research. Matching speakers whose speaking styles are similar is a new challenge. Now that the potential of the speaker descriptors is proven, many other speaker descriptors can be considered. For instance, it is known in literature that listener responses are usually placed around the end of a grammatical clause or sentence. Speaker descriptors describing the behavior around these moments may be helpful finding the right match.

Another interesting avenue for future research is in improving the development of the individual models included in the model collection. In the present study all individual models use the same features as input. However, since not every speaker uses the same cues to elicit listener response opportunities, not every feature will be helpful for each model. Feature selection for each individual model could potentially make the individual models stronger and in turn the speaker-adaptive model as a whole.

The presented speaker-adaptive model is a first step into the direction of modeling the mutual adaptation that takes place in interactions between interlocutors. In the current model only the variation and adaptation in speakers is considered, but listeners adapt their behavior as well. An interesting future direction we can take this approach is incorporating a listener profile as well. The MultiLis corpus offers two additional listeners in interaction with the same speaker. By learning individual models on these interactions as well and selecting a model from the model collection based on both a speaker profile and a listener profile could be the next step into modeling the mutual adaptation between interlocutors.

Another aspect we have not considered in the current study is that the speakers behavior is also dependent on other factors, such as his emotional situation and type of interaction. The model collection with speaker profiles presented in this paper could potentially be extended by including context, role and or emotional profiles. The main challenge for such models is to find an effective way to train good individual models for each combination of profiles for the model collection. To succeed either a lot of data or a succesful way to train individual models based on very limited training data is needed.

Acknowledgement

This material is based upon work partially 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.

7. REFERENCES

[1] *hCRF library*. http://sourceforge.net/projects/hcrf/.

[2] M. Argyle, R. Ingham, F. Alkema, and M. McCallin. The different functions of gaze. *Semiotica*, 7(1):19–32, 1973.

[3] J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[4] J. B. Bavelas, L. Coates, and T. Johnson. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580, 2002.

[5] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. *European ACL*, pages 51–58, 2003.

[6] T. L. Chartrand and J. A. Bargh. The chameleon effect: the perception-behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893–910, 1999.

[7] I. de Kok and D. Heylen. Appropriate and Inappropriate Timing of Listener Responses from Multiple Perspectives. In *Intelligent Virtual Agents*, pages 248–254, 2011.

[8] I. de Kok and D. Heylen. The MultiLis Corpus - Dealing with Individual Differences of Nonverbal Listening Behavior. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, pages 374–387. Springer Verlag, 2011.

[9] I. de Kok and D. Heylen. A survey on evaluation metrics for backchannel prediction models. In *Feedback Behaviors in Dialog*, pages 15–18, 2012.

[10] I. de Kok, D. Ozkan, D. Heylen, and L.-P. Morency. Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In *Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010.

[11] A. T. Dittmann and L. G. Llewellyn. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9(1):79–84, 1968.

[12] T. Drugman and A. Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Interspeech*, pages 1973–1976, 2011.

[13] C. Goodwin. *Conversational Organization: interaction between speakers and hearers*. Academic Press, 1981.

[14] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138, 2007.

[15] A. Gravano and J. Hirschberg. Backchannel-Inviting Cues in Task-Oriented Dialogue. In *Interspeech 2009*, pages 1019–1022, 2009.

[16] S. W. Gregory Jr. and B. R. Hoyt. Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psycholinguistic Research*, 11(1):35–46, Jan. 1982.

[17] L. Huang, L.-P. Morency, and J. Gratch. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In *Proceedings of Autonomous Agents and Multi-Agent Systems*, pages 1265–1272, 2010.

[18] M. Huijbregts. *Segmentation , Diarization and Speech Transcription : Surprise Data Unraveled*. Phd thesis, University of Twente, 2008.

[19] S.-H. Kang, J. Gratch, N. Wang, and J. H. Watt. Does the contingency of agents' nonverbal feedback affect users' social anxiety? In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, number Aamas, pages 120–127, 2008.

[20] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[21] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs. *Language and Speech*, 41(3-4):295–321, 1998.

[22] R. E. Kraut, S. H. Lewis, and L. W. Swezey. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4):718–731, 1982.

[23] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning*, pages 282–289, 2001.

[24] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *Intelligent Virtual Agents*, pages 25–36, 2005.

[25] L.-P. Morency, I. de Kok, and J. Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2011.

[26] Y. Okato, K. Kato, M. Kamamoto, and S. Itahashi. Insertion of interjectory response based on prosodic information. *Proceedings of IVTTA '96. Workshop on Interactive Voice Technology for Telecommunications Applications*, pages 85–88, 1996.

[27] D. Ozkan and L.-P. Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[28] C. Sun and L.-P. Morency. Dialogue act recognition using reweighted speaker adaptation. In *Proceedings of the 13th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL '12)*, pages 118–125, 2012.

[29] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[30] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.

[31] T. Watanabe and N. Yuuki. A Voice Reaction System with a Visualized Response Equivalent to Nodding. In *Proceedings of the third international conference on human-computer interaction, Vol.1 on Work with computers: organizational, management, stress and health aspects*, pages 396–403, 1989.