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

ABSTRACT
The goal of this paper is to analyze 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 identifies 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2129-7/13/12...
http://dx.doi.org/10.1145/2522848.2522860.

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 [17], 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 [16]. The absence of the expected listening behavior at such places can result in restarts (and often rephrases) from the speaker [14]. This affects the fluency of the conversation, which in turn affects speaker clarity and ultimately speaker comprehension [23, 3]. It has also been proven to hurt the rapport between interlocutors [15].

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], [16, 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, 26]. 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 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 profiles 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. [27]. 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. [25] presented the first multimodal approach. In their approach they combined Ward and Tsukahara’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 [20] and speech fluency [15]. The first multimodal machine learning approach was presented by Morency et al. [26]. 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. [11] recorded multiple listeners in interaction with the same speaker. Huang et al. [18] 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 [28] 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 successfully applied to adjust to speakers in the dialogue act recognition task [29].

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 speaking styles.

3. SPEAKER-ADAPTIVE PREDICTION MODEL OF LISTENER RESPONSES

In this section we introduce our speaker-adaptive prediction model of listener responses. We start with a general description of our prediction model and then we will explain the main novelty of our model in more details: speaker profiles designed to characterize the different speaking styles and select the proper prediction model. Finally, we describe how we build our collection of speaker-adaptive prediction models.

3.1 Overview

Our approach starts with an offline phase where we learn from an existing corpus of dyadic interaction individual prediction models of listener responses. These individual models are trained on only one specific interaction, meaning that each individual prediction model has a different speaker and/or listener. This allows to sample multiple speaking (and listening) styles. 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.

The next step is to define a comparative measure that allows to match similar speaking styles. To address this issue, we introduce the concept of speaker profile which is a computational description of the speaker behavior. We compute the speaker profile of all our individual models. More details about the speaker profile are presented in Section 3.2.

Once the offline data collection is performed, our approach is designed to automatically identify online the most relevant prediction model and predict listener behaviors. Figure 1 shows an overview of both the online and offline phases.
When a new speaker is interacting with our system, a speaker profile is computed to model his or her behaviors (i.e., speaking style). This speaker profile is compared to all speaker profiles in the model collection as depicted in the center of Figure 1 and nearest-neighbor based 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 selected prediction model is applied on the extracted input features from the new speaker giving us 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 key challenge with our speaker-adaptive prediction paradigm is how to identify similar speakers based on their behaviors (i.e., multimodal input features). Our challenge is not to find the exact same speaker among others, but finding a speaker with a similar speaking style that cues the moments where he/she expects a listener response in a similar way. This is a significant challenge since 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 analyzing techniques that are pooling all speaker and listener pairs together.

Features that are often found in conversation analysis literature to be associated with listener response opportunities include the pitch [22, 30, 16] and energy [22, 16] of the speech signal, pauses in speech [12, 5] and the eye gaze of the speaker [21, 2, 4]. Therefore, it is to be expected that differences are to be found in these same features. For instance, some speakers may use their gaze to cue the appropriate times for listener responses, while other may avert their gaze more than average. 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 usually represented as 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 the time when the feature is active and number of segments per minute.

To select a prediction model the speaker profile is compared to all speaker profiles in the model collection. There are many ways to compare two vectors and find the closest match. In this paper we performed nearest-neighbor using the Euclidean distance. This simple distance measured is shown empirically to succeed at improving prediction performance in our experiment section (see Section 5). We keep as future work the exploration of other distance measures.

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.
3.3 Model Collection Composition

As stated before the speaker profiles are used to select a the most similar prediction model from our collection which represents different speaking styles. A key step when building our model collection is to identify the top prediction models most helpful for the speaker adaptation procedure. The composition of the model collection is a balancing act between 1) the quality of the individual prediction models and 2) capturing the variability in speaking styles. In other words, the goal of the model collection is to have a good representation of the most predictive speaking styles.

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 our collection are selected based on their individual performance, while controlling for representation of the variability in speaking style.

4. EXPERIMENTS

The goals of our experiments are (1) to compare our speaker-adaptive approach with prior state-of-the-art approaches, and (2) to study the effect of different modalities in our speaker profiles.

This section will start with a description of the MultiLis corpus used for our evaluation. This will be followed by a detailed description of our learning technique for the individual prediction models. After this, the model selection of our user-adaptive learning approach will be described. Finally, the details of the evaluation methodology 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

Based on its prior success to predict listener backchannels, we selected Conditional Random Fields (CRF) [24] to learn the individual prediction models using the hCRF library [1]. CRF is a probabilistic discriminative model for sequential data labeling. CRF learns a mapping between a sequence of observations, in this case the input features describing the behavior of the speaker, and a sequence of ground truth labels. Our experiments were performed by selecting all positive samples of listener responses and the same amount of randomly selected moments where no listener response occurred as negative samples. The learned CRF model returns a prediction 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 these experiments, we compare our speaker adaptive approach with the CRF holistic approach which is learned from the whole dataset. For this comparison the following models were learned:

- **Holistic CRF Model**: Thirty-two holistic CRF models were learned. Each of these models was learned using 31 interactions from the MultiLis corpus as learning data and the remaining interaction as test data.
- **Individual Models**: Thirty-two individual models were learned. Each of these models was 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 were selected for the model collection of our speaker-adaptive multimodal prediction model (see Section 4.5).

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.

4.3 Input Features

All prediction models are learned on the input 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 is a visual feature. These features are:

- **Pitch**: The raw pitch values were extracted using the algorithm of Drugman and Alwan [13] 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 25 Hz.
- **Pitch Slope**: As a measurement of the pitch change overtime we compute the slope of the pitch by taking its first derivative.
- **Energy**: The energy of each speech frame is calculated on 32 ms Hanning windows with a shift of 10 ms and is expressed in dB.
- **Energy Slope**: As a measurement of the change in speech intensity, we compute the slope of the energy value by taking its first derivative.
• **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 [19]. 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.

### 4.4 Speaker Profiles

Our speaker-adaptive model is based on a collection of prediction models learned from multiple speaker-listener pairs. Each model is characterized by a speaker profile representing the behavior of the speaker. In this paper speaker profiles are defined by 10 speaker descriptors, each of them summarizing a behavior of the speaker over the course of the interaction. Our speaker descriptors are inspired from conversational analysis literature and include six acoustic features, two turn-taking features and two gaze features. These are:

- **Mean Pitch** - The mean pitch over the whole interaction calculated before z-score normalization (to model the differences between people).

- **Standard Deviation of Pitch** - The standard deviation of the pitch over the whole interaction also calculated using the raw pitch values before z-score normalization.

- **Mean Energy** - Mean of energy value over the whole interaction expressed in dB.

- **Standard Deviation of Energy** - Standard deviation of the energy values over the whole interaction.

- **Mean Energy Slope** - Mean of the energy slope values over the whole interaction.

- **Standard Deviation of Energy Slope** - Standard deviation of the energy slope values over the whole interaction.

- **Percentage of Speech** - The percentage of time voice is detected.

- **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.

During testing, all 10 descriptors are computed for the new speaker and compared with the speaker profiles found in the model collection. Our speaker-adaptive model selects the model whose speaker profile is the nearest neighbor match as measured by the euclidean distance.

### 4.5 Model Collection Composition

As previously stated the composition of the model collection is a balancing 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.

### 4.6 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 this threshold is determined during the validation phase [26, 11, 28]. 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. This is especially true for the individual models which are learned from only one interaction. 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. Recently, a dynamic thresholding method was proposed to deal with this issue in a more proper way [10].

Performance is measured using the averaged F1 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.
5. RESULTS

In this section we will present our experimental results starting in Section 5.1 with the comparison of our speaker-adaptative multimodal listener response prediction model with the state-of-the-art CRF model. Then Section 5.2 will present an analysis showing importance of our model collection composition. Finally, the importance of multimodality in our speaker profiles will be studied in Section 5.3.

5.1 Speaker-Adaptation

Figure 2 shows the comparison of our speaker-adaption approach (shown in red) with three baseline model, including the previously proposed CRF prediction model.

The performance of our speaker-adaptive model is 0.364 $F_1$ score (fourth bar in Figure 2). This is better than the performance of the state-of-the-art CRF model, which has a performance of 0.333 $F_1$ score (first bar in Figure 2). This difference is significant, $t(31) = 3.25, p = 0.001$.

If we only select the top 4 individual models

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 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 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 all these models proves that the speaker-adaptation accounts for most of the performance boost and not only the characteristics of the selected individual models.

5.2 Model Collection Composition

As previously stated, the composition of our model collection is a balancing act between 1) the quality of the individual prediction model and 2) capturing the variability in speaking styles. In this section, we analyze the characteristics of our model collection and compare different composition parameters.

To find the optimal model collection the number of models included in our model collection was varied from $N = 1$ to $N = 31$. For each of them we kept only the top $N$ individual models based on the mean performance as measured by the $F_1$ score. Figure 3 presents the results of varying the number of models in our collection (shown as a solid red line). The other lines are discussed in Section 5.3.

We observe that the maximum performance is achieved when the top 4 models are included in our collection with a peak of 0.364 $F_1$ score (right bar in Figure 2). The speaker-adaptive model that includes all individual models in its collection gives a performance of 0.323 $F_1$ score. This is worse than both the state-of-the-art CRF model and the best individual model. This suggests that the inclusion of some of the individual models hurts our performance.

One hypothesis is that 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 our model collection is to have a close match for any new speaker we may encounter. Including the models based on their performance might not be optimal since selected models could end up be close neighbors. To analyze this hypothesis, we performed a principal component analysis 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 this 2D space.
The results of this analysis are presented in Figure 4, where the four speakers selected for our collection are plotted in red and the remaining 28 models in black. The figure illustrates that the four selected models 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, we analyzed the importance of multimodality for our speaker profiles. A comparison was made between speaker profiles with multimodal speaker descriptors and unimodal speaker descriptors (acoustic, visual and turntaking). The comparison was made using our speaker-adaptive approach with varying only the speaker profiles. The results are presented in Figure 3.

Our speaker-adaptive prediction model with multimodal speaker descriptors is represented by the solid red line in Figure 3. For almost all model collection compositions the multimodal speaker descriptors outperform the unimodal speaker profiles. This analysis also shows that acoustic features may be best to define the speaker profiles (solid black line).

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a speaker-adaptive model for predicting listener responses. This speaker-adaptive model consists of a collection of individual prediction models that are trained on single interlocutor pairs. During the creation of our model collection we optimized the variability in speaker styles using the newly introduced concept of speaker profiles. When encountering a new speaker our approach compares the speaker profile of this new speaker to all the speaker profiles in our collection. The closest matching speaker is used to predict listener response opportunities for the new speaker.

When compared to a state-of-the-art CRF model our approach showed a statistically improvement over a previous proposed approach. The performance of our approach 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]). Our 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.

Our speaker-adaptive approach 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 from literature that listener responses are usually placed around the end of a grammatical clause or sentence [14]. Using speaker descriptors only around these moments may be helpful in finding better matches.

Another interesting avenue for future research is to improve the performance of the individual models included in the model collection. In our study all individual models used 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 a specific model. Feature selection for each individual model could potentially improve performance.

Another aspect we have not considered in the current study is that speakers behaviors are also dependent on other factors, such as emotional state, relation between interlocutors and topic. Our speaker profiles could potentially be extended by including context, role and/or emotional profiles. Our major challenge will be to find sufficient interactions to model all combinations of profiles. Researchers may find a way to train individual models based on very limited training data is needed, enabling such a study.

Our speaker-adaptive model is a first step into the direction of modeling the mutual adaptation that takes place during dyadic interactions. In our model, only the variability and adaptation to speaking styles is considered, ignoring the potential difference in listening styles. An interesting future direction would be to incorporate a listener profile as well. The MultiLis corpus offers two additional listeners for each speaker. By using these additional interactions and selecting models based on both the speaker profile and the listener profile could be the next step into modeling the mutual adaptation between interlocutors.

Acknowledgement

This research has been supported by the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 231287 (SSPNet).

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


