Learning and Evaluating Response Prediction Models using Parallel Listener Consensus

Iwan de Kok
Human Media Interaction,
University of Twente
koki@ewi.utwente.nl

Derya Ozkan
Institute for Creative
Technologies, University of
Southern California
ozkan@ict.usc.edu

Dirk Heylen
Human Media Interaction,
University of Twente
heylen@ewi.utwente.nl

Louis-Philippe Morency
Institute for Creative
Technologies, University of
Southern California
morency@ict.usc.edu

ABSTRACT

Traditionally listener response prediction models are learned from pre-recorded dyadic interactions. Because of individual differences in behavior, these recordings do not capture the complete ground truth. Where the recorded listener did not respond to an opportunity provided by the speaker, another listener would have responded or vice versa. In this paper, we introduce the concept of *parallel listener consensus* where the listener responses from multiple parallel interactions are combined to better capture differences and similarities between individuals. We show how parallel listener consensus can be used for both learning and evaluating probabilistic prediction models of listener responses. To improve the learning performance, the parallel consensus helps identifying better negative samples and reduces outliers in the positive samples. We propose a new error measurement called $F_{consensus}$ which exploits the parallel consensus to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. We present a series of experiments using the MultiLis Corpus where three listeners were tricked into believing that they had a one-on-one conversation with a speaker, while in fact they were recorded in parallel in interaction with the same speaker. In this paper we show that using parallel listener consensus can improve learning performance and represent better evaluation criteria for predictive models.

Categories and Subject Descriptors

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

ICMI-MLMI'10 , November 8-12, 2010, Beijing, China
Copyright 2010 ACM 978-1-4503-0414-6/10/11 ...$10.00.

General Terms

Algorithms, Human Factors, Theory

1. INTRODUCTION

During a conversation the interlocutors are in constant interaction with each other. Speakers check to what extent the listeners are paying attention to the conversation and whether they have understood the message. The listener signals this by a wide arrange of responses, commonly called backchannels. These responses have been proven to improve both the quality of the narrative of the speaker and the understanding of the listener [1, 14]. Therefore, it is regarded as one of the important aspects of human behavior which should be modelled, to create engaging virtual humans. Several systems (e.g. [8, 13, 19]) already include such models and it has been shown to improve engagement and speaker fluency of the user of the system [8].

When working towards a model for listener responses in an interactive virtual human system, development and evaluation of such a model is usually done on the basis of a corpus of recorded human-human interactions. The assumption is that a model which can reproduce the listeners recorded in a corpus most accurately is the best model. However, as there are a lot of individual differences between listeners, one listener might respond to certain cues from a speaker, whereas another listener would not have or vice versa. So part of the data will be mislabelled as negative samples. This is a problem in both development and the evaluation of the model. In development the sequential probabilistic model will include overlaps, since some of what should have been positive samples are labeled as negative samples. During evaluation the model may produce responses at times the recorded listener did not respond, but another listener would have. Such responses should not be considered as a false positive, where in fact they will.

In this paper, we introduce the concept of parallel listener consensus for learning and evaluating probabilistic prediction models of listener responses (see Figure 1). To improve the learning performance, the parallel consensus helps identifying better negative samples and reduces outliers in the positive samples. We propose a new error measurement called $F_{consensus}$ which takes advantage of the parallel con-

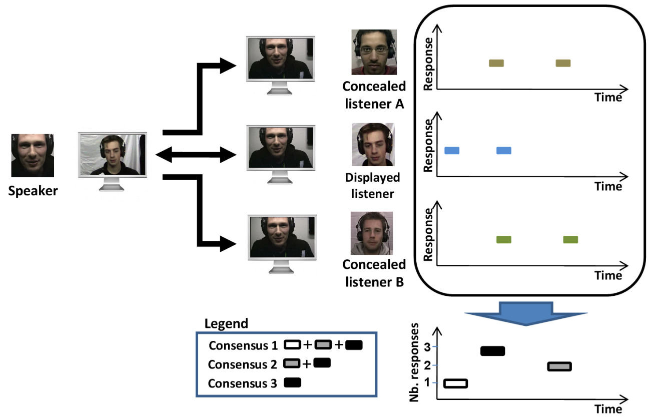

Figure 1: Parallel Listener Consensus. Parallel Listener Consensus is defined as the combination of the listener responses from multiple parallel interactions. The parallel listener consensus captures the differences and similarities between individuals. In this paper, we show how parallel listener consensus can be used for both learning and evaluating probabilistic prediction models of listener responses.

sensus to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. We present a series of experiments using the MultiLis Corpus [4] where three listeners were tricked into believing that they had a one-on-one conversation with a speaker, while in fact they were recorded in parallel in interaction with the same speaker. In this paper we show that using parallel listener consensus can improve learning performance and represent better evaluation criteria for predictive models.

In Section 2 we discuss previous work on listener response prediction models and their methods to cope with individual differences. In Section 3 we present the parallel listener consensus and how it can be used to learn and evaluate prediction models of listener responses. The experimental setup is introduced in Section 4 and results are discussed in Section 5.

2. RELATED WORK

Over the past decade, several researchers have developed models to predict listener responses. Ward and Tsukahara [20] created a handcrafted rule-based model using low pitch and utterance length as cues for listener responses. When analyzing the performance of their predictive rule they conclude that 44% of the incorrect predictions were cases where a listener response could naturally have appeared, as judged by one of the authors, but in the corpus there was silence or, more rarely, the start of a turn.

Cathcart et al. [3] identified the same problem as well. In their shallow model of backchannel continuers based on pause duration and n-gram part-of-speech tags they remark that human listener differ markedly in their own backchanneling behavior and pass up opportunities to provide a backchannel. Their attempt to deal with this is testing their model on high backchannel rate data, reasoning that the more backchannels an individual produces, the fewer opportunities they are likely to have passed up.

Both of these approaches only offer a solution for the problem during evaluation, but not for the problem during development. Noguchi and Den [17] do offer such a solution. A machine learning approach is taken for modelling backchannel behaviors based on prosodic features. This approach requires a collection of positive and negative examples of appropriate context for a listener response to occur. These examples were built by collecting listener responses from participants in a study, in which the participants were asked to hit the space bar on a keyboard at times where they thought a listener response was appropiate while watching recorded stimuli of a speaker. The stimuli were several pause-bounded phrases and constitute a single conversational move (on average 2.91 phrases per stimuli). Each stimuli was shown to 9 participants. By counting the number of participants that responded to a phrase positively, each phrase is classified either as an appropriate context for a listener response, or an inappropriate context for a listener response, or indecisive.

A similar approach was used by Huang et al. [11]. They called this Parasocial Consensus Sampling (PCS). They let observers indicate appropriate listener response moments for whole conversations instead of conversational moves. An interesting result from the study by Huang et al. [11] is that the best listener response model need not be the model that reproduces the recorded listener in a corpus most accurately. From the results of the PCS they animated a virtual listening agent based on the displayed listener and another one on the consensus of the observers. They let new observers watch an interaction between the original speaker video and the animated agent. The agent based on the consensus of the observers was perceived as more believable and said to show more rapport than the agent based on the displayed listener. Both offer no evaluation of these observation based acquisition of listener responses. It remains to be seen whether people actually respond at the moments they indicated during PCS when placed in the same interaction.

To our knowledge no previous work has proposed a method of learning prediction models of listener responses based on multiple parallel interactions. In this paper, we propose two new techniques for learning and evaluating response prediction models using parallel listener consensus.

3. PARALLEL LISTENER CONSENSUS

Parallel Listener Consensus can be defined as the combination of the listener responses from multiple parallel interactions. The parallel listener consensus captures the differences and similarities between individuals. In the following section we will explain the advantage this concept brings to the prediction of listener responses in more detail. First the MultiLis Corpus will be introduced in Section 3.1. Then we will explain how we combine the recordings of multiple listeners into consensus instances. How these consensus instances can be used to improve the state of the art of listener response prediction in both learning and evaluation will be discussed in Sections 3.3 and 3.4 respectively.

3.1 Parallel Listener Corpus

The MultiLis corpus [4] is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totalling 131 minutes. Participants (29 male, 3 female, mean age 25) were assigned the role of either speaker or listener during an interaction. The speakers summarized a video they has just seen or reproduced a recipe they has just studied for 10

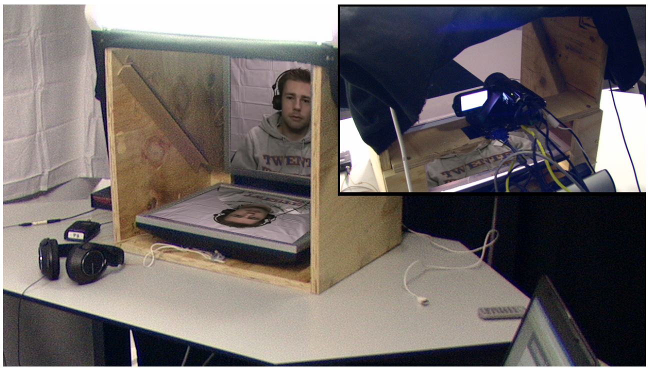

Figure 2: Picture of the cubicle in which the participants were seated. It illustrates the interrogation mirror and the placement of the camera behind it which ensures eye contact.

minutes. Listeners were instructed to memorize as much as possible about what the speaker was telling. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener.

What is unique about this corpus is the fact that it contains recordings of three individual listeners to the same speaker in parallel, while each of the listeners believed to be the sole listener. The speakers saw one of the listeners, believing that they had a one-on-one conversation. We will refer to this listener, which can be seen by the speaker, as *displayed listener*. The other two listeners, which can not be seen by the speaker, will be refered to as *concealed listeners*. All listeners were placed in a cubicle and saw the speaker on the screen in front of them. The camera was placed behind an interrogation mirror, positioned directly behind the position on which the interlocutor was projected (see Figure 2). This made it possible to create the illusion of eye contact. To ensure the illusion of a one-on-one conversation was not broken, interaction between participants was limited. Speakers and listeners were instructed not to ask for clarifications or to elicit explicit feedback from each other.

The recordings were annotated manually for a number of features. For the listener the corpus includes annotations of head, eyebrow and mouth movements, and speech transcriptions. What we refer to as a listener response can be any combination of these various behaviors, for instance, a head nod accompanied by a smile, raised eyebrows accompanied by a smile or the vocalization of "uh-huh", occurring at about the same time. For each of these responses we have marked the so-called *onset* (start time). The onset of a listener response is either the stroke of a head movement, the start of a vocalization, the start of eyebrow movement or the start of a mouth movement. When different behaviors combine into one listener response, either the head movement or vocalization was chosen as onset (whichever came first). If there was no head movement or vocalization present, either the eyebrow or mouth movement was chosen as onset (whichever came first).

During this annotation all different kind responses of the listener are annotated. We use the 2456 responses from the MultiLis corpus with a head movement and/or a vocalization as our ground truth labels (from a total of 2798 including the smiles and eye brow responses). Having ground truth labels as homogeneous as possible is a desirable property while learning a prediction model in order to model the cues provided by the speaker as accurately as possible. To create more homogeneous ground truth labels, we excluded the responses with only a smile or eye brow movements. These responses are closer to the concept of what Goodwin [7] refers to as assessments (smiles are usually responses to funny content or situations and eye brow movements to surprising (raise) or confusing (frown) content), while the responses with a head movement and/or a vocalization are closer to the concept of backchannel continuers [18]. Since we are not including a representation of the content in our feature set, we will be unable to model the conditions to which the assessment responses are a reaction.

Algorithm 1 Response consensus building algorithm

```
Require: sorted $allResponses$ from all Listeners
Require: $consensus\_window$
  while $allResponses$ is not empty do
    $firstResponse$ = earliest in $allResponses$
    $tStart$ = start time of $firstResponse$
    $thisConsensus$ = all responses starting in $(tStart + consensus\_window)$
    $lastResponse$ = latest in $thisConsensus$
    $tEnd$ = start time of $thisLastResponse$
    $allConsensus = allConsensus + [tStart, tEnd]$
    $allResponses = allResponses - thisConsensus$
  end while
  return $allConsensus$
```

3.2 Building Response Consensus

In this section, we present our algorithm to build listener responses consensus. While the consensus algorithm easily scale to any number of listeners, for simplicity we present the algorithm for three listeners.

Our corpus contains the recordings of three individual listeners in interaction with the same speaker. To create our model we need to establish the consensus which we can use for learning and evaluating our model. Therefore we need consensus responses from the three listeners to appear at the same time. But what is 'at the same time'? When do different listeners respond to the same cues from the same speaker, and how large is the "window of opportunity" to start a response to a cue?

Our parallel listener consensus is based on the observation that the window of opportunity is correlated with the gap between two responses from the same listener. In other words, the minimum time between two responses from the same listener gives us a bound for the fusion of parallel listeners. By analyzing the listener interactions in our training corpus, we found the minimal response gap to be 714ms. To ensure that our algorithm does not group two responses from the same listener, the *consensus window* is set to 700 ms.

In Algorithm 1 the algorithm is presented. A forward looking search is performed. When an hitherto unassigned response is encountered, the algorithm checks whether there are more responses which start within the consensus window of 700 ms from the start time of this response. If there are, all of these are grouped together with the response. The

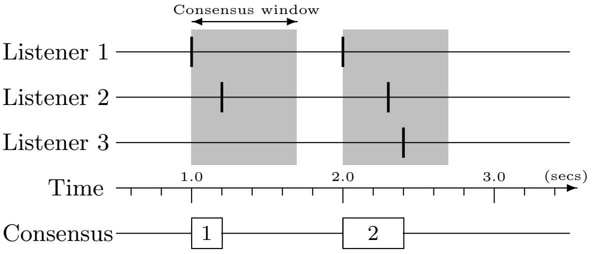

Figure 3: Example of the consensus building algorithm. At time 1.0s the algorithm has encountered a listener response from listener 1. It checks whether there are more responses from other listeners within the *consensus window* of 700ms. There is a response from listener 2 at time 1.2s, thus these are grouped into consensus instance 1, which starts at time 1.0s and ends at time 1.2s. The algorithm continues with the next unassigned response and repeats the process and creates a consensus instance from 2.0s to 2.4s in the same way, by combining the three responses from listener 1, 2 and 3.

start time of the consensus instance is the onset of the first response the end time of the consensus instance is the onset of the latest response included in the consensus. This thus corresponds to the ”‘Window-of-Opportunity”’ as found in the data that starts with the beginning of the first listener response and ends with the beginning of last listener response in the consensus. Note that this means that if a consensus of only one response is created the start and end time of the consensus are identical. After a consensus is created we continue our forward looking search for the next unassigned response.

In Figure 3 an example is given of the consensus building algorithm. Using our corpus, the algorithm created 1733 consensus instances. There are 1140 consensus 1 instances (instances contain only one response), 465 consensus 2 instances (two responses) and 128 consensus 3 instances (responses from all three listeners).

3.3 Using Consensus during Learning

The goal of our prediction model is to create real-time predictions of listener reponses based on features of human speakers (see Figure 4). For this purpose, we use a machine learning approach in which we train a sequential probabilistic model from a database of consensus interactions and use this trained model to generate listener responses. A sequential probabilistic model takes as input a sequence of observation features (e.g., the speaker features) and returns a sequence of probabilities (i.e., probability of listener response). During learning, the ground truth labels which mark the appropriate response opportunities are required as well as the speaker features. How this ground truth labels are established is what differentiates our approach from traditional methods.

We have established several consensus for listener responses in the corpus. By only considering the displayed listener we can regard this corpus as any other corpus of recorded one-to-one interactions with its limitations caused by individual differences in listening behavior. Some people provide a lot of listener responses, while others use their responses more sparsely. When the recorded listener provides only a few responses during the interaction, it does not necessarily mean that there are only a few response opportunities. Thus, some of the data is mislabelled as negative samples, while in fact these would be valid response opportunities.

But we also have the listening behavior of the concealed listener at our disposal and can utilize this during learning. By recording more people less of our data is mislabelled as negative examples. Using the displayed listeners gave in total 879 responses to response opportunities, but consensus building identified 1733 response opportunities. This is almost a doubling of the number of positive samples.

The parallel consensus does not only improve the quality of negative samples and increase the number of positive samples, but also provides information about the importance of each response opportunity, reducing the effect of outliers. To some response opportunities all three listeners respond; whereas to some others, only two or mostly one of the listeners responds. The response opportunities to which three listeners respond are more clearly cued by the speaker than response opportunities to which two listeners respond and even more than opportunities to which only one listener respond. The speaker will expect a response at these moments. By emphasizing these response opportunities during learning your model should be more tuned to predict these response opportunities and will therefore result in a better model.

3.4 Using Consensus during Evaluation

The consensus gives us a more reliable performance measure during evaluation. We propose a new evaluation criteria based on multiple listener consensus which exploits the parallel consensus to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. The basis of our consensus-based measure is the F_1 measure, which is the weighted harmonic mean of precision and recall. Precision is the number of correctly predicted listener responses divided by the total number of predicted listener responses (correct or not). It is a measure of exactness, highlighting the effect of false positives (i.e., predicted responses mislabeled as positive). Recall is the number of correctly predicted listener responses divided by the total number of listener responses (i.e., ground truth). It is a measure of completeness, highlighting the effect of false negatives (i.e. listener responses that were not predicted correctly). The main idea behind our new consensus-based measure is that precision and recall should not be computed using the same ground truth elements. We introduce the concepts of Consensus Exactness and Completeness:

CONSENSUS EXACTNESS The typical approach for computing the false positives (necessary for the precision measurement) is to look at the ground truth responses from the displayed listener. The problem with this approach is that while the displayed listener may not have given a listener response at a specific point in time, another person would have given a response at that moment. With the multiple listener consensus framework we have the listening behaviors from concealed listeners at our disposal to counter this shortcoming. We propose that consensus exactness should take into account all listeners and classify a prediction as false positive only if none of the listeners responded at that moment. This concept implies that precision should be computed using all consensus instances as

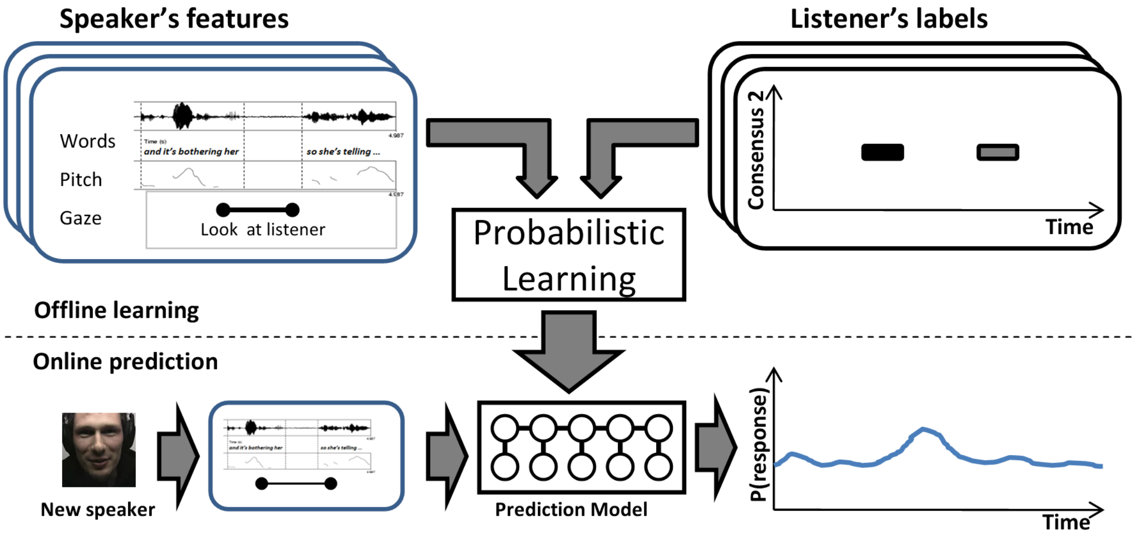

Figure 4: Learning using parallel listener consensus. A sequential probabilistic model is trained offline using as input a sequence of observation features (e.g., the speaker features) and the ground truth labels from the consensus data (Consensus 2 in this figure). The prediction model returns a sequence of probabilities (i.e., probability of listener response) during the online testing.

ground truth.

CONSENSUS COMPLETENESS If all consensus instances were used to compute the false negatives (necessary for the recall measurement), the perfect model would be a model which is able to predict all response opportunities from any listener. This model would end up giving responses at a much higher frequency then any individual person. The experiment of Huang et al. [11] has shown that a virtual human based on the consensus of several listeners is perceived as most believable when the rate of generated responses is similar to the average rate of all listeners. Based on this observation, we propose that consensus completeness should be correlated with a consensus level which has an average number of ground truth responses equal to the average rate from all listeners.

Based on these two concepts, we define our consensus-based evaluation criteria $F_{consensus}$ as follows:

$$F_{Consensus} = 2 * \frac{Precision_{all} * Recall_t}{Precision_{all} + Recall_t}$$

where the precision $Precision_{all}$ is calculated using all consensus ground truth responses and the recall $Recall_t$ is calculated using only the ground truth responses from the consensus t (i.e., at least t listeners responded at that moment). t is automatically selected such that the average rate of ground truth responses is as close as possible to the desired rate (average rate from all listeners). In our experiments, the average rate was 6.3 responses per minute. The closest match was Consensus $t = 2$ (i.e., at least two listeners responded at that moment) with 4.5 responses per minute. The combination of $Precision_{all}$, which takes care of the mislabelled negative samples, and the $Recall_t$, which keeps an average response rate, results in a more reliable performance measurement.

4. EXPERIMENTAL SETUP

In this section, we first describe the machine learning technique we used to create our prediction model and the methodology for evaluating it. Then we explain the five strategies we used used in our experiments for ground truth and the speaker features used as input.

4.1 Learning Prediction Model

In our experiments, we use Conditional Random Fields (CRF) [15], which is a probabilistic discriminative model for sequential data labeling. A CRF learns a mapping between a sequence of observations and a sequence of labels. Every gesture class has a corresponding state label. During evaluation, we compute marginal probabilities for each state label and each frame of the sequence using belief propagation. The optimal label for a specific frame is chosen to be the label with the highest marginal probability. Applying a threshold on the marginal probability of the gesture, we assign a positive label to a frame if the marginal probability was larger than the threshold. We use the hCRF library [10] for the training of our CRF models.

Testing is performed on an hold-out set of 10 randomly selected interactions. The remaining 21 dyadic interactions were used for learning. All models evaluated in this paper were trained with the same training set and tested on the same test set. The test set does not contain individuals from the training set. Validation of model parameters was performed using a 3-fold strategy on the training set. The objective function of the CRF model contains a regularization term to prevent overfitting. During training and validation, this regularization term was validated with values 10^k, for $k = -3..3$.

In all models the ground truth labels are normalized to the same length of 700ms. The mean start time of the responses included in each consensus instance is calculated. Each instance starts at the 350ms before this mean start time and ends 350ms after it.

4.2 Prediction Models

As discussed in Section 3.3 the MultiLis Corpus provides opportunities to improve the prediction of listener responses, since it includes recordings of parallel listeners. In this section we will describe the five different models which we trained. In each model we used a different strategy to combine the ground truth of the parallel listeners.

DISPLAYED LISTENER ONLY Our first model consists of a CRF chain model trained with using responses of only the displayed listener as the ground truth labels. This model is our main baseline for our experiments since most previous work used this approach (such as [3, 16, 20]). We refer to this model as the *DL only* model in the rest of the paper.

ALL LISTENERS In the second model, we use responses of both the primary listener and the two secondary listeners in the same session. For the secondary listeners, we duplicate speaker-listener pairs by using the same speaker for both listeners. These duplicated listener-speaker pairs can be seen as different sessions in which the speaker has the same features and listeners have their own responses. We refer to this model as *ALL* model in the rest of the paper.

CONSENSUS 1, 2 AND 3 The last three models implement our consensus building strategy described in Section 3.2. The *Consensus 1* model includes all consensus instances. So all the response opportunities to which at least one listener (either the displayed listener or one of the concealed listeners) has responded are used as ground truth label. The *Consensus 2* model only includes response opportunities to which at least two listeners have responded as ground truth label and the *Consensus 3* model only opportunities to which all three listeners have responded.

4.3 Multimodal Features

Previous research has identified several cues the speaker gives to elicit a listener response. We extracted the following features: lexical features (see, for instance [16] for evidence of these features as cue), pause (see [3, 16]), gaze (see [2, 5, 16]) and prosodic features (see [9, 17, 20]).

LEXICAL AND PAUSES The lexical features were extracted using the Dutch automatic speech recognition software SHoUT [12]. We collect the recognized words with their start and end times and the start and end times from silences. From these results we created utterance and pause features. These features are mutual exclusive, where utterance segments are defined as interpausal units, where the minimum length of the silence between two utterances is 100 ms. These segments of silence longer than 100 ms are defined as pauses.

EYEGAZE AND BLINK Eyegaze and blink features were manually annotated. For eyegaze the human coder annotated whether the speaker was looking at the listener (directly into the camera) or not. Gazes at the listener were occasionally interrupted by blinks of the speaker. Even though the gaze was interrupted for a moment, the listener would still have the perception that the speaker is addressing him/her. Therefore we created the "continued gaze" feature where the blinks between and after a gaze annotation are included into the interval. From both the normal gaze and the continued gaze features we created a "blinked" variant, which only includes the gaze intervals which were preceeded by a blink.

Model	F_1	Precision	Recall
Baseline (DL Only)	**0.265**	0.268	0.262
All Listeners	0.255	0.188	0.392
Consensus 1	0.225	0.166	0.352
Consensus 2	**0.264**	0.199	0.391
Consensus 3	0.239	0.170	0.402

Table 1: The performance of our five models measured using only the displayed listeners ground truth labels.

	Consensus 1	Consensus 2
Model	F_1	F_1
Baseline (DL Only)	0.278	0.253
All Listeners	**0.377**	0.255
Consensus 1	0.318	0.213
Consensus 2	**0.375**	**0.287**
Consensus 3	0.364	0.256

Table 2: The performance of our five models measured Consensus 1 and Consensus 2 ground truth labels.

PROSODY For the extraction of prosodic features we used openSMILE [6] to extract the pitch (F0) value at a 10ms interval. To remove some non-valid values where the pitch was not computed for short periods of time, we apply a smoothing filter (size=5) on the whole pitch data of each speaker. Then these values are discretized into percentiles.

5. RESULTS AND DISCUSSION

During our experiments we have trained five models using various strategies to establish ground truth. We evaluated these models on four different performance measures. In the following sections we will discuss the various results, starting with the perfomance on $F_{consensus}$ (Table 3), then on displayed listener only (Table 1) and finally on Consensus 1 and 2 (Table 2).

DISPLAYED LISTENER ONLY As a baseline we measured the performance of our response prediction model on the Displayed Listener Only (DL Only). Table 1 shows the performance of our five models on this measure. Our result with learning on DL Only ($F_1 = 0.265$) on this case (our baseline model) is comparable to the result of Morency et al. [16] ($F_1 = 0.256$) but on a different corpus. Looking at the other approaches which the MultiLis corpus allowed us to take, we can see that learning on Consensus 2 achieves comparable performance ($F_1 = 0.264$) and also the performance of the ALL model is only slightly worse. The other approaches perform not as good as the traditional approach of using DL Only for learning.

CONSENSUS 1 AND 2 As discussed in Section 3.4 this corpus provides us with more information than only the responses of the displayed listener. We also have the responses

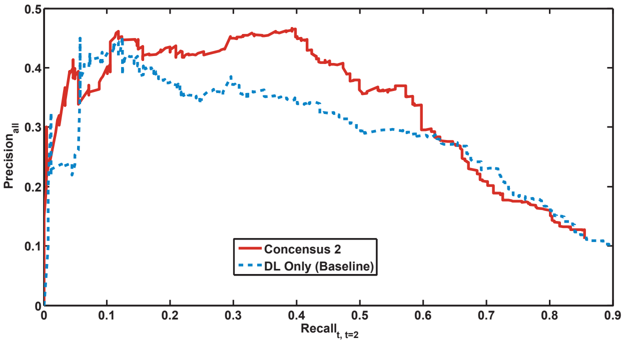

Figure 5: Precision and recall graph for the Displayed Listener Only (Baseline) model (red striped line) and the Consensus 2 model (blue line). With most of the thresholds, including the thresholds giving the highest $F_{consensus}$ score, the Consensus 2 model outperforms the Displayed Listener Only (Baseline) model.

Model	$F_{consensus}$	Precision$_{all}$	Recall$_t$
Baseline (DL Only)	0.347	0.419	0.297
All Listeners	0.425	0.370	0.499
Consensus 1	0.358	0.311	0.421
Consensus 2	**0.439**	0.373	0.534
Consensus 3	0.417	0.338	0.542

Table 3: The performance of our five models measured on our $F_{consensus}$ measure. The difference between our Baseline model and Consensus 2 is marginally significant, $p = 0.054$.

of the concealed listeners available to us and this information can also be used during evaluation to get a more precise performance measure dealing with exactness and completeness. Since the production of a listener response is optional, we see in our corpus that the displayed listener and the concealed listeners do not always respond at the same time. The displayed listener may miss response opportunities, to which one or both of the concealed listeners did respond. A prediction of our model at such a missed response opportunity should not be a wrong prediction, since according to our corpus, these are moments where listeners do provide responses. In our corpus Consensus 1 provides the broadest coverage of these moments. Therefore, we also looked at the performance of our models using Consensus 1 as ground truth (see Table 2). On this measure the All Listeners ($F_1 = 0.377$), Consensus 2 ($F_1 = 0.375$) and Consensus 3 ($F_1 = 0.364$) models perform significantly better than the Displayed Listener Only ($F_1 = 0.278$) model. The Consensus 1 ($F_1 = 0.318$) model has a performance in between the other models.

However, Huang et al. [11] have shown that this does not result in the most believable and attentive virtual human. This is because your response rate is too high if you generate a response on all predicted response opportunities. They have shown that a virtual human which responds at moments most people would respond is the most believable. In our corpus these are the moments were two or three listeners responded to the same opportunity at the same time (Consensus 2). Using these ground truth labels the response rate is closest to the response rate of the average listener. Again, the Consensus 2 model ($F_1 = 0.287$) performs best on this measure, but the differences with the Displayed Listener Only ($F_1 = 0.253$), All Listeners ($F_1 = 0.255$) and Consensus 3 ($F_1 = 0.256$) models are not significant.

F-CONSENSUS Measuring the performance on Consensus 2 re-introduces the problem of the Displayed Listener Only evaluation, where response opportunities are mislabelled as negatives. Our $F_{consensus}$ measure solves the problems of exactness (mislabels) and completeness (missed labels) by calculating precision on Consensus 1 and recall on Consensus 2. The results of our model on this measure are presented in Figure 5 and Table 3. The Consensus 2 model ($F_{consensus} = 0.439$ performs better than the DL Only model ($F_{consensus} = 0.347$). The difference is marginally significant, $p = 0.054$. Also the models trained on ALL ($F_{consensus} = 0.425$) and on Consensus 3 ($F_{consensus} = 0.417$) perform better.

So, overall learning on Consensus 2 performs best in all cases, which proves the use of parallel listener consensus in the learning phase. Furthermore using $F_{consensus}$ as performance measure gives us a more reliable performance measure which takes advantage of the parallel consensus to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. Especially on this measure the advantages of using parallel listener consensus shows in the learning phase.

6. CONCLUSION

In this paper, we introduced the concept of *parallel listener consensus* where the listener responses from multiple parallel interactions are combined to better capture differences and similarities between individuals. We showed how parallel listener consensus can be used for both learning and evaluating probabilistic prediction models of listener responses. To improve the learning performance, the parallel consensus helps identifying better negative samples and reduces outliers in the positive samples. Across all metrics learning on the Consensus 2 ground truth labels performed best.

Furthermore, we proposed a new performance measure called $F_{consensus}$ which takes advantage of the parallel consensus to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. We presented a series of experiments using the MultiLis Corpus where three listeners were tricked into believing that they had a one-on-one conversation with a speaker, while in fact they were recorded in parallel in interaction with the same speaker. We showed that using parallel listener consensus can improve learning performance and represent a better evaluation criteria for predictive models.

At this time we only used three parallel recorded listeners, but getting more listeners in parallel and more samples in general should improve the performance of these techniques even more. More listeners would mean having an even bigger coverage of the response opportunities and therefore less false negative samples. Furthermore, a better threshold would be achieved for the minimum consensus agreement on the positive samples (for both ground truth labels in learning and in the $F_{consensus}$ measure), reducing outliers.

A broader application of the proposed techniques is on sequential annotated data with low annotation agreement between annotators, especially if the cause of the low agreement is the fact that it is hard to recognize the behavior in the sequence (as opposed to classify it with the correct label). The structure of that data is very similar to the data we worked with. There are several ground truths for the same sequence and some annotators may have missed moments which other annotators have noticed. A prediction at the time only one annotator made an annotation might not actually be wrong, it is maybe so subtle that only that annotator noticed it. On the other hand high agreement moments should definitely not be missed by your predictor. Our techniques are designed to deal with data with exactly these characteristics.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE), 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] J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[2] 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.

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

[4] I. de Kok and D. Heylen. The MultiLis Corpus - Dealing with Individual Differences of Nonverbal Listening Behavior. *Towards Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, 2010.

[5] S. Duncan Jr. On the structure of speaker-auditor interaction during speaking turns. *Language in society*, 3(2):161–180, december 1974.

[6] F. Eyben, M. Wöllmer, and B. Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, pages 576–581, Amsterdam, Netherlands, 2009.

[7] C. Goodwin and M. H. Goodwin. Concurrent Operations on Talk: Notes on the Interactive Organization of Assessments. *IPRA Papers in Pragmatics*, 1(1):1–54, 1987.

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

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

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

[11] 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*, Toronto, Canada, 2010.

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

[13] S. Kopp, J. Allwood, K. Grammer, E. Ahlsén, and T. Stocksmeier. Modeling Embodied Feedback with Virtual Humans. *Modeling Communication with Robots and Virtual Humans*, pages 18–37, 2008.

[14] 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.

[15] 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.

[16] 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, 2010.

[17] H. Noguchi and Y. Den. Prosody-based detection of the context of backchannel responses. In *Fifth International Conference on Spoken Language Processing*, 1998.

[18] E. A. Schegloff. *Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences*, pages 71–93. Georgetown University Press, Washington, 1982.

[19] M. Schröder, E. Bevacqua, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wöllmer. A demonstration of audiovisual sensitive artificial listeners. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–2, Amsterdam, Netherlands, september 2009.

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