# LISTENINGHEADS.

IWAN DE KOK

# LISTENING HEADS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
Prof. dr. H. Brinksma
on account of the decision of the graduation committee
to be publicly defended
on Thursday, September $12^{th}$, 2013 at 16:45

by

Iwan Adrianus de Kok

born on November $28^{th}$, 1982
in Dongen, The Netherlands

**Composition of the Graduation Committee**:

| | | |
|---|---|---|
| Prof. Dr. Ir. | A.J. Mouthaan | Universiteit Twente |
| Prof. Dr. | D.K.J. Heylen | Universiteit Twente |
| Prof. Dr. Ir. | A. Nijholt | Universiteit Twente |
| Prof. Dr. | F.M.G. de Jong | Universiteit Twente |
| Prof. Dr. Ir. | M. Pantic | Universiteit Twente and Imperial College |
| Prof. Dr. | H. Bunt | Tilburg University |
| Dr. | J. Edlund | KTH Royal Instutute of Technology |
| Dr. Ir. | L.-P. Morency | USC Institute for Creative Technologies |

The research reported in this thesis has been carried out at the Human Media Interaction (HMI) research group of the University of Twente.

CTIT Ph.D. Thesis Series No. 13-266
Centre for Telematics and Information Technology
P.O. Box 217, 7500 AE
Enschede, The Netherlands.

SIKS Dissertation Series No. 2013-29
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-90-365-0648-9
ISSN: 1381-3617 (CTIT Ph.D. Thesis Series No. 13-266)
DOI: 10.3990/1.9789036506489
http://dx.doi.org/10.3990/1.9789036506489

Typeset with LaTeX. Printed by Ipskamp Drukkers B.V., Enschede.
Cover design: Iwan de Kok

Voor mijn ouders

This thesis has been approved by:

Prof. Dr. D.K.J. Heylen
Prof. Dr. Ir. A. Nijholt

# Acknowledgements

Vier jaar onderzoek samengevat. Dat ligt nu voor je. Een persoonlijke mijlpaal, maar zeker geen mijlpaal die ik alleen heb bereikt. Vele mensen hebben mij direct of indirect gesteund de afgelopen vier jaar en die verdienen een bedankje.

Als eerste wil ik mijn ouders bedanken. Jullie hebben me altijd gesteund in alles wat ik wilde doen en daar ben ik jullie erg dankbaar voor. Zonder jullie steun en de gegeven vrijheid zou ik nooit zo ver gekomen zijn. Ik laat het te weinig merken, maar ik hou van jullie.

Ik wil ook mijn flatgenoten bedanken. Ook na elf jaar is Flat Lodewijck nog altijd een fijne plek om naar terug te keren na een werkdag. Er is altijd wel iemand met een luisterend oor beschikbaar. Door de jaren heen hebben veel mensen deel uit gemaakt van onze flat, maar de ongedwongen, relaxte Lodewijck-sfeer is altijd gebleven, één van de redenen dat ik er ook tijdens mijn promotie-traject ben blijven wonen.

Vooral wil ik Dirk bedanken voor het geloof in mij en de vrijheid die ik van je kreeg om mijn eigen route te kiezen. De gesprekken, niet alleen over het onderzoek, maar ook over de alledaagse dingen waren erg prettig. Ik herinner me vooral nog het weekend in Parijs, waar je Ronald, Dennis, Khiet en mij uitgenodigd had om je appartement te bezoeken en ons langs allerlei winkeltjes met lekkernijen leidde.

Ook wil ik alle andere collegas van HMI bedanken voor de leuke, leerzame en gezellige tijd die ik er gehad heb. Een aantal collegas wil ik nog specifiek noemen. Ronald; voor zijn altijd aanwezige enthousiasme, die leven in de brouwerij brengt. De vele conferenties/summerschools die we samen bezochten en de vakantiedagen die we er in IJsland en Californi aan vast plakten waren ook altijd erg gezellig. De samenwerkingen aan een aantal papers en discussies over onderzoek in het algemeen waren ook erg leerzaam. Khiet, voor het delen van ons kantoor op de vakgroep, Lynn, voor het doorlezen en verbeteren van mijn thesis, Hendri voor alle hulp met het creëren van het MultiLis corpus, Charlotte en Alice voor de administratieve hulp bij de dagelijkse gang van zaken. Ook alle bachelor-referaat studenten die meegeholpen hebben met onderdelen van het onderzoek en alle deelnemers aan de onderzoeken wil ik bedanken.

Furthermore, I would like to thank Louis-Philippe Morency. In 2008 I went off to the USC Institute of Creative Technologies for an internship, thinking I would be working with Jonathan Gratch on “something” with virtual humans - the subject was not determined beforehand. Instead I was working with a crazy, enthusiastic French Canadian guy who introduced me to listener behavior prediction, the topic I would be working on ever since. This first research experience also convinced me to pursue

a career in research. Since then we continued to work together from time to time, which I really enjoy. Your push for perfection is inspiring. I'm glad you could be part of my graduation committee. I would also like to thank the other committee members for participating in my defense.

De mensen van A.D.S.K.V. Slagvaardig wil ook bedanken voor het geven van een heerlijke uitlaatklep elke week op het knotsbal veld en de gezellige activiteiten er om heen. Ook naar de pokermiddagen met mijn vrienden van de middelbare school, de D&D weekenden met vrienden van de universiteit en barbecues met mijn ouddoegroepgenoten kijk ik altijd weer uit en ik ben blij dat we deze tradities nog steeds in stand weten te houden.

# Contents

# 1

# Introduction

One of the issues to address in research on spoken dialogue systems and embodied conversational agents is to take care that the system produces appropriate behavior when the person interacting with the system is speaking. In human-human conversations, listeners produce feedback to the speaker in the form of nods, facial expressions, and short expressions such as 'uh huh', 'mmh' for instance. We all know from experience that in the absence of such signals - which we will refer to as listener responses - communication problems can arise.

One solution to the issue would be to have the system produce listener responses at random. However, it will turn out in this case that there are moments during the speech in which listener responses are expected by the speaker which are not being produced by the system and moments where the system produces a response which is somewhat unexpected. A designer of an artificial listener would like to avoid misplaced responses. The research described in this thesis addresses precisely this issue. It proposes and evaluates algorithms to have a system produce appropriate listener responses of a certain kind. These are based on studies - also part of this thesis - of what real human listeners do when they interact with another human speaker. For this part, special methods were introduced that try to take into account the fact that not all listeners behave in the same way. While there are cases where a listener response is expected and cases where they are highly unexpected, there are also many moments during a conversation where producing a response is fine, but not producing it is fine as well - where one person would produce a response and another person would not.

## 1.1 Listening Behavior

Having a conversation requires complex coordination between verbal and nonverbal behavior to shape the information which is passed on from one interlocutor to the other. This is true for the interlocutor who is speaking as well as for the interlocutor currently listening. The speaker provides the information, while the listener is constantly providing feedback to the speaker. Researchers have found that the func-

tion of this feedback is to signal contact, perception, understanding and/or other attitudinal reactions [128, 37, 3, 53, 36, 14] to the speaker. Feedback is regarded as an important aspect of the grounding process between interlocutors [37, 18, 36]. In this grounding process the common ground in terms of mutual knowledge, mutual beliefs and mutual assumptions is established. Before advancing further into the dialog the interlocutors make sure that everyone has a clear understanding of what has come before. The (absence of) feedback plays an important role in this aspect. The speaker uses the (absence of) feedback from the listener to assess the current understanding of the common ground by the listener and adapts to the listener's needs if needed [44, 55, 58, 56, 59, 53, 7, 14]. For example, when the listener gives a signal of misunderstanding in reaction to vital information or does not acknowledge it, the speaker can choose to repeat, rephrase or give more details about this information. Alternatively, when the listener gives a clear signal of understanding early on in an explanation the speaker can choose to shorten this explanation. These improvements in quality of the speaker's speech have been shown by several researchers [97, 92, 91, 93, 7] as well as the subsequent improvement in understanding of the speaker's speech by the listener [93, 7, 19, 50, 158]. Furthermore, listening behavior has been proven to increase the rapport between interlocutors [27, 62].

Throughout the literature, many names are give for the instances of feedback, such as backchannel (activity/feedback) [160, 71, 153], minimal response [126, 48], reactive token [33], accompaniment signal [86], acknowledgment token [83, 43, 161, 84], aizuchi [104, 87], and many more. In this thesis the term listener responses will be adopted for these behaviors [40, 7, 8, 52]. Many of the aforementioned terms refer to a subset of the behaviors with a specific function. Since in this thesis no analysis will be made into the function of the behaviors, the neutral term listener response is preferred.

The form of these listener responses ranges from short vocal utterances such as "uh-huh", "yeah" or "okay" to various head gestures, smiles [21] and other facial displays. Researchers have analyzed the acoustic characteristics and semantics of the vocal listener responses [65, 76, 25, 29, 134, 10] and the various forms of head gestures [66, 67, 82, 2, 30, 72, 119] and facial displays [105] used as listener response. In this thesis the main focus will be on head gestures, in particular head nods. Most of the listener responses in the corpus recorded for this thesis are head nods and generation experiment will be conducted with head nods as well.

Bavelas *et al*. [7] make the distinction between generic and specific listener responses. In this distinction the generic listener responses are not specifically connected to what the speaker is saying. One could easily interchange one generic listener response with another and both would be equally appropriate. The main function of these generic listener responses is to signal attendence and a general notion of understanding to let the speaker know he/she can continue. Typical generic listener responses include nods and minimal verbal utterance such as "mm-hmm" or "yeah". The focus of the thesis is on this generic type of listener responses, as the head nods and vocalizations in the corpus recorded for this thesis are all of the this type.

The specific listener responses in this distinction are tightly connected to what the speaker is saying. These listener responses give an assessment of what has been

said and as such cannot always be interchanged. Typical specific listener responses include emotional facial displays, such as smiles, fear or surprise, and short verbal utterances such as “oh, wow!” or “that’s sad”. These specific listener responses give an assessment of what has been said and/or a signal of understanding for specific parts of what has been said, shown by repetitions or additions. Goodwin [57] makes a similar distinction between continuers and assessments.

Differences in listening behavior in terms of the number of listener responses given and the form of the listener responses have been found between interlocutors of different gender [75, 103, 96, 126, 41, 109, 47] and culture [157, 104, 33, 156, 135, 47, 69, 159]. However, even when these factors are the same the listening behavior of individuals is seldom the same. Giving a listener response is often optional. An interaction will not immediately break down if one or even a few opportunities to give a listener response are passed up by the listener. Little is known about the factors that determine whether an opportunity can be passed up by the listener or not. One of the goals of this thesis is to give more insight into these factors by analyzing what will be called response opportunities (see Section 1.3) in human-human interactions.

## 1.2 Embodied Conversational Agents

Besides studying listening behavior in interactions between humans, the behavior has also received attention from the virtual agents and robotics communities. The qualities that appropriate listening behavior brings to an interaction, such as improvement of speaker’s speech and increased rapport between interlocutors, are highly desired qualities for the applications for systems such as companion or information giving agents.

Many aspects of listening behavior for embodied conversational agents have been investigated. Some researchers have focussed on the perception of generated listener responses [60, 130, 145, 15, 73, 143, 17, 120, 122]. Other researchers have focussed on the detection of visual listener responses [106] or vocal listener responses [112, 111]. Furthermore, researchers have worked on the interpretation of and adaptation to listening behavior [22, 23]. Finally, researchers have also investigated the effect of generated listening behavior on the user of the embodied conversational agent system [61, 62, 132, 144, 127, 146, 147, 125].

The focus of this thesis and the remainder of this section will be on computational models that generate listening behavior in response to the speaker. These models analyze the speaker and generate listener responses at the appropriate times to signal attention to, understanding and/or assessment of what has been said. The distinction made by Bavelas *et al.* [7] between generic and specific listener responses has also been adapted by many models for listening behavior for embodied conversational agents. Researchers have realized that the generation of these types of listener responses each require a different approach. For the generation of generic listener responses researchers have adopted the development of reactive models, while for the generation of specific listener responses deliberate models are developed. Furthermore, the reactive models use more shallow feature, whereas the deliberate model use more semantic features. Since listening behavior includes both types of responses

both approaches ultimately need to be merged and to coordinate with each other.

In this section some examples of both types of models for embodied conversational agents are presented.

### 1.2.1 Reactive Models

Reactive models for listening behavior are focussed on the actions of the speakers. Actions of the speaker directly determine whether a listener response is generated or not. These models analyze features extracted from the audio and video signals that record the behavior of the speaker. In these features the reactive models are looking for patterns in the behavior of the speaker that are associated with listener responses. The patterns these models are looking for are based on observations in corpora of human-human interactions. The models are either handcrafted based on results of conversational analyses or automatically learned with machine learning techniques.

The listening behavior of the Gandalf system [138] reacts to pauses from the speaker. After a pause with a duration of 110 ms the system generates a verbal or nonverbal listener response. Similar to the Gandalf system, the REA system [26] uses pauses to detect suitable moments for listener responses.

Maatman *et al*. [102] also generated the listening behavior of the Rapport agent with a reactive model. The model consists of a set of rules found in literature that directly match behavioral patterns of the speaker to reactive behavior from the listener. Head nods are generated in reaction to lowered pitch and raised loudness in the speech signal. When a disfluency is detected in the speech signal a posture shift, gaze shift and/or frown is generated. Furthermore, postures, gaze and head gestures of the speaker are mimicked. In a second version of the Rapport agent [80] the handcrafted rules are replaced by data-driven models.

Many more reactive models have been developed besides the ones that are applied in an embodied conversational agent. These listener response prediction models are developed and evaluated on corpora of example human-human interaction. A more comprehensive overview of work on such reactive listener response prediction models will be presented in Chapter 6. The computational models of listening behavior developed in this thesis are also reactive listener response prediction models.

### 1.2.2 Deliberative Models

Specific listener responses can be seen as co-telling acts in which the listener gets involved in the conversation by showing their attitude towards, assessment of or (short) contribution to what has been said by the speaker. For automatic generation of such listener responses for embodied conversational agents more elaborate internal representations about emotions and attitudes may be required. The agents needs to recognize and interpret what has been said and form an attitude towards this. For this behavior detecting patterns in the speaker's actions may not suffice: reasoning about the internal state of the listener may be required as well. Deliberative models for specific responses include this reasoning. Many of the deliberative models include a reactive component as well.

Kopp *et al*. [90] proposed a deliberative model for listening behavior for the

embodied conversational agent Max that works with textual user input. The model reasons about five concepts to determine the listening behavior. These five concepts are Contact, Perception, Understanding, Acceptance and Emotion/Attitude. Contact represents whether the embodied conversational agent is still in contact with the user. Perception is run on a word-by-word basis and evaluates whether the system knows each word. Understanding represents whether the user input can be succesfully interpreted. Acceptance evaluates whether the user input complies with the agent's current beliefs, desires or intentions. Emotion/Attitude represents the emotional reaction as appraised by the emotion system of the embodied conversational agent. A probabilistic rule-based system is in place that reacts to events triggered by changes of these five concepts.

Bevacqua *et al*. [16] proposed a deliberative model for the Sensitive Artificial Listener agent. Besides the verbal and nonverbal behavior of the speaker the model generated backchannels based on the speaker's interest level and the mental state of the agent. This mental state describes the attitude of the agent towards the interaction. In the SAL agent four mental states are defined. These four mental states are angry/argumentative, gloomy, happy and sensitive/pragmatic.

Wang *et al*. [148] proposed a deliberative model for listening behavior that is capable of changing its behavior based on the current conversational role of the agent. The model discriminates the listener roles of addressee, side-participant, eavesdropper and overhearer. The model combines the incoming signals of the speaker with the current role and goals of the listener to determine the listening behavior. The agent is capable of expressing understanding, (mirror) emotion, thinking behavior and role switching behavior, such as expressing intentions to enter or leave the conversation. All these require incremental understanding and assessment of the speaker.

## 1.3 Contribution

In this section an overview of the contributions will be given. This thesis is a methodological exploration of individual differences and similarities in listening behavior and how these differences and similarities can be used in the development and evaluation of listener response prediction models for embodied conversational agents.

The contributions of the thesis can summarized by the following.

- **The Concept of Response Opportunities** - Central to the work presented in this thesis is the concept of *response opportunities*. A response opportunity can be defined as a window in time where a listener response is appropriate. The concept is similar to the concept of *turn-transition relevance places* [129], a concept known in conversation analysis for places where it is relevant for another interlocutor to take the turn. Recently, Heldner *et al*. [70] referred to the concept of response opportunities as *backchannel relevance space*.

- **Methods for Collecting Multiple Perspectives on Listening Behavior** - One of our goal is to be able to recognize these response opportunities in an interaction. For this we need an annotated data set of human-human recordings in which all response opportunities in the interactions are known. But since giving a listener

response at a response opportunity is optional, no listener will respond to all response opportunities. This means that not all response opportunities in an interaction are identified by looking at the listener responses of a single listener, which is usually what is recorded. To discover all response opportunities in an interaction the listener responses of multiple listeners are necessary. In this thesis we will explore and compare different methods to collect these listener responses of multiple listeners. In this thesis we will call a collection of listener responses of a single individual a *perspective*.

- **Conversational Analysis of the Characteristics of Response Opportunities** - To be able to recognize response opportunities we will analyze the actions of the speaker in the seconds before a response opportunity. Since we identified each response opportunity by combining the multiple perspectives from the previous contribution, we also know how many listeners responded to each response opportunity. This gives us a measure for the graded optionality of the response opportunities - we assume that there are various degrees to how optional it is to give or not to give a listener responses. In the analysis we look into which actions of the speaker causes this *graded optionality* or in other words what makes some response opportunities more compelling to respond to than others.

- **Methods for Learning Listener Response Prediction Models using Differences and Similarities between Interlocutors** - The ultimate goal of the presented research is to build a computational model to generate listening behavior for an embodied conversational agent. The focus of the thesis towards this goal is on reactive *listener response prediction models*. These models look for behavioral patterns in the speaker's actions and relate these to the likelihood of a listener response. Methods are explored that use the multiple perspectives to develop more accurate listener response prediction models and performance measures for these models.

## 1.4 Overview

Development of computational models of listening behavior for embodied conversational agents typically follows a series of steps. As a first step recordings of human-human conversations are collected. The purpose for these recordings is to analyze the listening behavior. The results of these analyses give pointers to the interplay between behaviors of the speaker and the decision of the listener to give a listener response or not. The analyses give insight into which behaviors of the speaker are important to measure as input features to be able to generate the appropriate listening behavior. Based on this knowledge and using the recordings as data a listener response prediction model can be learned. Finally, the prediction model can be evaluated either by comparing the predictions of the model to the listening behavior in the recordings or by subjective evaluation of the generated behavior by human observers.

The structure of the thesis will follow along this development cycle for listener response prediction models. In Part I the methods of data collection will be introduced. Part II will cover the conversational analyses performed on the recorded corpus. In

Part III the prediction models will be learned and evaluated. The thesis will be concluded in Part IV reflecting on the thesis and looking ahead to future work.

### 1.4.1 Data Collection

Part I of the thesis will present the data collection methods introduced in this thesis. The goal of the data collection is to collect multiple perspectives on listening behavior to identify the response opportunities in an interaction.

In Chapter 2 the Parallel Recording method will be introduced. With this method the MultiLis corpus was recorded which will be used throughout the thesis. In this method three listeners are recorded in interaction with the same speaker. Due to the setup of the recording only one of the listeners can be seen by the speaker, but all believe they are addressee in the interaction. By recording three listeners in parallel three perspectives of appropriate listening behavior are collected. By combining these perspectives response opportunities can be identified. These response opportunities are moments in time where a listener response can be given. By looking at the number of listeners that responded to a response opportunity we can look at the graded optionality of response opportunities.

Chapter 3 will confirm that recording three listeners still does not give a complete coverage of all the response opportunities in the interaction. The Parasocial Sampling method will be presented as a method to collect even more perspectives of appropriate listening behavior. In the parasocial sampling method first introduced by Huang *et al*. [79] subjects watch recorded speakers and give listener responses as if they were interacting with the speaker. These parasocial listener responses are recorded on the keyboard. Results will be presented validating this data collection method as a substitute for actual recordings for the purpose of collecting the timing of listener responses. Combining the collected perspectives with the parallel recorded perspectives increases the coverage of all the response opportunities in the interaction and further diversifies them into important and less important response opportunities.

Beside the parasocial sampling method Chapter 3 will also introduce the Individual Perceptual Evaluation method. In this method generated listener responses are individually evaluated on their appropriateness. Subjects observe interactions between a recorded speaker and a virtual listener. The subjects are tasked with judging each individual listener response from the virtual listener on appropriateness. When the subjects judge a listener response to be inappropriate they hit a key on the keyboard. This evaluation method can be used to collect perspectives on inappropriate listening behavior as well as evaluating the performance of a virtual listener on the level of individual behaviors instead of on a general impression.

### 1.4.2 Conversational Analysis

Part II of the thesis will present the conversational analysis results on the collected perspectives on appropriate and inappropriate listening behavior. The goal of the conversational analysis is to analyze the relationship between the behavior of the speaker and presence or absence of a response opportunity.

Before starting the actual conversational analyses a manipulation check will be

performed in Chapter 4 to confirm that there are no significant differences in the behavior of the displayed listener that can be seen by the speaker and the two concealed listener that cannot be seen. The manipulation check will show that the closing of the interaction loop between the speaker and the two concealed listeners did not change their behavior significantly. Thus, for the conversational analyses the listeners can be regarded as equal.

In Chapter 5 the results of the conversational analyses will be presented. In the conversational analyses the behavior of the speaker around the response opportunities collected in Part I will be analyzed. The conversational analysis starts with a qualitative study looking at the content of the speaker's speech in the vicinity of response opportunities and inappropriate moments for listener responses. Observations will show relations to sentence structure (listener responses before (part of) the rheme is completed are considered inappropriate), conversational structure (listener responses in reaction to a summarizing or refining statement are more appropriate) and proximity of earlier responses (producing two similar listener responses in close succession is considered inappropriate).

Analysis of the speech activity of the speaker shows that response opportunities are placed near or right after the end of an utterance. This will be illustrated by an analysis of the presence or absence of speech in the vicinity of response opportunities and the energy of the speech signal. Furthermore, results will be presented that show that the pitch of the speech either falls or rises to low and high values, respectively, starting 750 ms before the response opportunity. Finally, results will be presented that show that the speaker specifically looks at the listener at response opportunities. All these cues for response opportunities are found to be more frequent at response opportunities where more than one listener responded.

### 1.4.3 Learning Prediction Models

Part III of the thesis will present three listener response prediction models learned on the MultiLis corpus. The experiments focus on finding methods to use the different perspectives on listening behavior and differences in speaking behavior to develop more accurate and adaptive listener responses prediction models and more accurate evaluation metrics. Each models focusses on improving different aspects of the learning process. The first two models will focus both on acquiring more accurate ground truth samples. The first model will focus on acquiring more accurate positive samples and the second model on acquiring more accurate negative samples. The third model focusses on better handling the learning data by splitting the data and learning multiple models that each represent a different speaking style.

Before presenting the three listener response prediction models an overview of the state-of-the-art will be presented in Chapter 6. The survey will focus on highlighting differences between approaches with regard to the corpus the model is learned on, the features that are used as input for the models, the modeling technique that is used and the method of evaluation.

The listener response prediction model that will be presented in Chapter 7 focusses on selecting better positive samples for the ground truth labels. The parallel recording method identified many response opportunities each with one, two or three listeners

that responded to them. The experiments will show that using only the response opportunities with responses from two or more listeners as positive ground truth labels performs better than alternative selections. Furthermore, a new evaluation measure will be introduced which values correctly predicting response opportunities with a response from the majority of the listeners more highly, while not ignoring the response opportunities where only a minority of the listeners responded.

The listener response prediction model that will be presented in Chapter 8 focusses on selecting better negative samples for the ground truth labels. The presented approach will use the perspectives of inappropriate moments for listener responses collected using the Individual Perceptual Evaluation method. The model is iteratively learned. After each iteration the model is evaluated using the Individual Perceptual Evaluation method. The collected inappropriate moments from this evaluation are used in a following iteration as negative samples of the ground truth for learning. The results will show that the listening behavior generated by this prediction model is more appropriate according to the human observers.

The listener response prediction model that will be presented in Chapter 9 focusses on adapting to the speaker. The focus of this approach is on acknowledging that people differ from each other and that one listener prediction model will probably not work for every speaker. Speakers have their own voice characteristics, fluency of speech and other behavior patterns. Because of these differences there are differences in the way response opportunities are cued by these speakers. The presented speaker-adaptive model will learn individual models for different speakers. When encountering a new speaker the model will analyze the characteristics of the speaker and compare those to the characteristics of the speakers it has a model for. The model that is learned on the closest matching speaker is selected. Results will show that this speaker-adaptation results in a significant improvement in performance.

In Chapter 10, the final chapter of Part III, a method will be presented to integrate these listener response prediction models into an embodied conversational agent. Based on the time since the last generated listener response, the proposed dynamic thresholding method varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to adapt the behavior to the situation, targeted role or personality of the virtual agent. Results will show that the generated behavior is more stable under changing conditions than the behavior of the traditional fixed threshold.

So, to conclude, this thesis will combine work in the areas of data collection, conversational analysis and machine learning to develop computational models for the generation of listening behavior for embodied conversational agents. The contributions to each of these areas will focuss on capturing, analyzing and modeling the similarities and differences in listening behavior that exist between individuals. Part IV will reflect on these contributions and will discuss directions for future work.

# Part I

# Collection of Multiple Perspectives

# 2

# Parallel Recording

A listener response is usually not given at the listener's whim. These responses are signals from the listener towards the speaker that the information has been received, understood and potentially evaluated by the listener. They are tied to actions of the speaker as part of the grounding process that takes place between interlocutors. In an interaction there are specific moments where the listener can give a response, namely the moments where the speaker provides the listener with information that needs to be grounded. We call the moment a response can be given by a listener a *response opportunity*.

The data that will be presented in this chapter, involves conversations in which a speaker was matched with several listeners where everyone was made to believe they were in a one-to-one conversation. There are moments in this data where only one of the listeners produces a listener response. This indicates that listeners do not need to produce a listener response at every response opportunity. [1]

However, a typical listener will not provide a listener response at each response opportunity. These moments are *opportunities* and there is no fixed rule that states that a response is required. This characteristic optionality of listening behavior brings a challenge in building a computational model of said behavior. This causes variation in the type, timing and number of listener responses between individuals. One passed up opportunity for a listener response will not immediately break the interaction, but a total absence of responses will. The question is, at which moments is it essential to respond as a listener and which moments can be passed up.

In order to gain insight into which response opportunities can be passed up and which require a response, we need multiple perspectives on appropriate listening behavior in response to certain speaker actions. One example of a listener will not give us complete coverage of all response opportunities in the interaction.

Typically a corpus only has one perspective on appropriate listening behavior recorded and covers only the response opportunities to which the recorded listener

[1] Of course taking a more individual, cognitive perspective, one could say that what counts as a response opportunity for one listener may not count as a response opportunity for another. The point of view adapted in this thesis is that a response opportunity in the data is a moment which at least one of the listeners (or perspectives) regards as an opportunity to give a listener response.

has responded to. Examples of such corpora on which listener response analyses have been conducted include the HCRC Map Task Corpus [4, 28], the CID Corpus [13, 12, 11] and the Rapport Corpus [62, 108]. In these corpora one example of appropriate listening behavior is recorded in response to the actions of the speaker. However, another individual placed in the same interaction will most likely not act in exactly the same way. This listener will most likely respond to partially the same response opportunities and partially other ones. Even if the response is to the same opportunity, it can take a different form.

One could argue that corpora that feature multi-party conversations, such as the AMI Corpus [24, 74], includes multiple examples of appropriate listening behavior. Oftentimes one of the participants is speaker, while the other three are listening. However, frequently the floor [68] of the interaction organizes such that the speaker is addressing one of the participants, while the other two overhear this (short) interaction between the two. The behavior of an addressee is different than that of an overhearer [54, 35, 98, 49]. The speaker expects responses from the addressee, while none are expected from the overhearer. Thus, an addressee is more likely to respond than an overhearer. Even if the speaker is addressing all three of the remaining participants, the speaker can only look at one of them at a time, so conditions are not exactly the same for all three. While the responses from the overhearer can be used to identify response opportunities passed up by the addressee, the differences between addressees and overhearers make an analysis of the graded optionality of the response opportunities flawed.

To overcome this, we recorded a corpus where the conditions are exactly the same for all listeners. All three listeners perceive themselves to be the addressee of the interaction. The three perspectives on appropriate listening behavior in this corpus will give us a more complete coverage of all response opportunities in the interaction.

In the remainder of the chapter the corpus will be introduced in more detail. In Section 2.1 the setup for the data collection will be explained. In Section 2.2 the details about the recordings and annotations will be presented.

## 2.1 The Data Collection

The goal of this data collection was to record multiple perspectives of appropriate listening behavior in response to actions of a speaker. Therefore, we need recordings of multiple listeners in response to the same speaker and trying to ensure that the reactions of these listeners are as natural as possible. We, therefore, needed to create the illusion that the listeners believed that they were the only listeners in the interaction and thereby the addressee of the speaker. Once this illusion would be broken people may change their behavior pattern from addressee to overhearer. An effect of this might be a lower response rate, since an overhearer does not need to give responses to the speaker, since the speaker does not expect them to.

In this corpus we aimed to record interactions between one speaker and three listeners. To achieve this, without the participants realizing this fact, the interactions were video-mediated. The listeners were made to believe they are having a one-on-one conversation with the speaker. Also the speaker was unaware of the special setup,

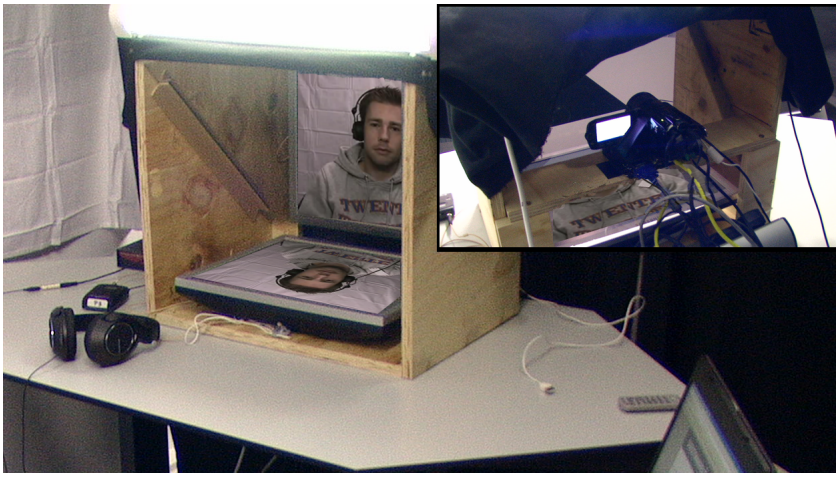

**Figure 2.1:** Picture of the cubicle in which each participant was seated. It illustrates the interrogation mirror and the placement of the camera behind it which ensures eye contact.

seeing only one of the listeners.

The data collection designed to record the corpus is presented in more detail in the following sections. The setup of the data collection will be discussed in Section 2.1.1. The procedure during recording will be discussed in Section 2.1.2, the tasks of the participants are discussed in Section 2.1.3 and the extra data that we collected such as the demographics, personality of participants will be presented in Section 2.1.4.

### 2.1.1 Setup

Each of the participants sat in a separate cubicle. The digital camcorders which recorded the interaction, were placed behind a one-way mirror onto which the interlocutor was projected (see Figure 2.1). This ensured that the participants got the illusion of eye contact with their interlocutor. In Figure 2.2 one can see that the listeners appear to be looking into the camera which was behind the mirror. This video was also what the participants saw during the interaction. All participants wore a headphone through which they could hear their interlocutor. The microphone was placed at the bottom of the autocue set up and was connected to the camcorder for recording.

During the interaction speakers were shown one of the listeners (the *displayed listener*) and could not see the other two listeners (the *concealed listeners*). All three listeners saw the recording of the same speaker and all three believed that they were the only one involved in a one-to-one interaction with that speaker. Distribution of the different audio and video signals was done with a Magenta Mondo Matrix III, which is a UTP switchboard for HD-video, stereo audio and serial signals. Participants remained in the same cubicle during the whole experiment. The Magenta Mondo Matrix III enabled us to switch between distributions remotely.

### 2.1.2 Procedure

In total eight sessions were recorded. For each session there were four participants invited (in total there were 29 male and 3 female participants, with a mean age of 25). At each session, four interactions were recorded. The participants were told that in each interaction they would have a one-on-one conversation with one other participant and that they would either be a speaker or a listener. However, during each interaction only one participant was assigned the role of speaker and the other three were assigned the role of listeners. Within a session, every participant was a speaker in one interaction, was once a displayed listener and appeared twice as concealed listener.

In order to be able to create this illusion of one-on-one conversations we needed to limit the interactivity of the conversation, because as soon as the displayed listener would ask a question or start speaking, the concealed listeners would notice this in the behavior of the speaker and the illusion would be broken. Therefore the listeners were instructed not to ask questions or take over the role of speaker in any other way. However we did encourage them to provide short feedback to the speaker.

### 2.1.3 Tasks

The participants were given tasks. The participants that were given the role of speaker during an interaction either had to retell the events from a video clip or give the instruction for a cooking recipe. The listeners had the task to remember as much as possible.

For the retelling of the video speakers were instructed to watch the video carefully. For the data collection the 1950 Warner Bros. Tweety and Sylvester cartoon "Canary Row"[2] and the 1998 animated short "More" by Mark Osborne[3] were used. The speaker had to remember and tell as many details as possible, since the listener would be asked questions about the video after the interaction. To give the speakers an idea of the questions which were going to be asked, they received a set of 8 open questions before watching the video. After watching the video they had to give the questions back so that they would not have anything to distract them.

After the retelling both the speaker and the listeners filled out a questionnaire with 16 multiple choice questions about the video. Each question had four alternative answers plus the option "I do not know" and for the listener the extra option "The speaker did not tell this".

For the second task the speaker was given 10 minutes to study a cooking recipe. As stimuli a tea smoked salmon recipe and a mushroom risotto recipe were used. After the interaction both the listener and the speaker needed to reproduce the recipe as completely as possible in the questionnaire afterwards. As performance measure the reproduction of the recipe by the listeners was scored. Two points could be scored for the title and the number of persons the recipe was intended for; for the items on the ingredient list 23 points; for the description of the procedure 25 points; for a maximum total of 50 points.

[2]Canary Row (1950): http://www.imdb.com/title/tt0042304/
[3]More (1998): http://www.imdb.com/title/tt0188913/

To control for differences in the quality of the summary of the video or reciting of the recipe between interactions, the three listeners were ranked among themselves. The listener with the best score received a 1, the second best a 2 and the third best a 3. Ex aequo listeners received the same ranking.

### 2.1.4 Measures

Before the recordings we asked participants to fill out their age and gender and we had them fill out personality and mood questionnaires. For personality we used the validated Dutch translation of the 44 item version of the Big Five Inventory [85]. For mood we used seven out of eleven subscales from the Positive and Negative Affect Schedule - Expanded Form (PANAS-X, 41 items) [155] and the two general positive and negative affect scales. Furthermore we used the Profile of Mood States for Adults (POMS-A, 24 items) [137]. For both PANAS-X and POMS-A we used unvalidated Dutch translations made by the authors. Participants were instructed to assess their mood of "today".

After each interaction speakers filled out the Inventory of Conversational Satisfaction (ICS, 16 items) [157], questions about their task performance (5 items) and questions about their goals during the interaction (3 items). The listeners filled out an adapted version of the rapport measure [62] with additional questions from the ICS (10 items in total, e.g. "There was a connection between the speaker and me."). Some questions of the 16 items ICS relate to talking, which the listener did not do in our experiment, so they were left out. Furthermore the listeners answered six questions about the task performance of the speaker, such as "The speaker was entertaining" or "The speaker was interested in what he told". All questions were 5-point Likert Scale.

After the complete session, when all four interactions were finished subjects were debriefed and were asked which interaction they preferred; whether they had believed the illusion of always having one-on-one interaction, and if not, at which moment they had noticed this; in which interaction they thought the speaker could see them; about the delay of the mediated communication, audio and video quality (3 items).

## 2.2 The MultiLis Corpus

The main motivation for doing the experiment was to collect the recordings of the interactions. In this section the more details about the resulting recordings and the collected annotations from the recordings are discussed.

### 2.2.1 Data

In total 32 interactions were recorded (8 for each task), totalling 131 minutes of data (mean length of 4:06 minutes). All the interactions were in Dutch.

Audio and video for each participant was recorded in synchrony by the digital camcorders. Synchronisation of the four different sources was done by identifying the time of a loud noise which was made during recording and could be heard on all audio signals.

**Figure 2.2:** Screenshot of a combined video of the four participants in an interaction.

Videos are available in high quality (1024x576, 25fps, FFDS compression) and low quality (640x360, 25fps, XviD compression). Audio files are available in high quality (48kHz sampling rate) and low quality (16kHz sampling rate). Furthermore a combined video (1280x720, 25fps, XviD compression) of all four participants in a conversation is available (for a screenshot, see Figure 2.2).

### 2.2.2 Annotations

Speakers were annotated on eye gaze and smiles. Listeners were annotated on head, eyebrow and mouth movements and any speech they produced was transcribed as well. For this annotation we used the ELAN annotation tool [20].

For the listeners the annotations were made in a three step process. First the interesting regions with listener responses were identified. This was done by looking at the video of the listener with sound of the speaker and marking moments in which a response of the listener to the speaker was noticed. In the second step these regions were annotated more precisely on head, brows and mouth movements. Speech of the listener was also transcribed by hand. In the third and final step the onset of the response was determined.

In the following subsections the annotation scheme for each modality will be explained in more detail. In each annotation scheme left and right are defined from the perspective of the annotator.

**EYE GAZE** Annotation of the speakers' gaze provides information about whether they were looking into the camera (and therefore looking at the listener) or not and whether there was blinking. For each of these two features a binary tier was created. Annotations were done by two annotators who each annotated half of the sessions. One session was annotated by both. Agreement (calculated by overlap / duration ) for gaze was 0.88 and for blink 0.66.

**HEAD** For listeners the shape of the head movements. An annotation scheme of 12

categories was developed. The 12 categories and the number of annotations in each category are given below. Several movements had a lingering variant. Lingering head movements are movements with one clear stroke followed with a few more strokes that clearly decrease in intensity. If during this lingering phase the intensity or frequency of the movement increases again a new annotation is started. In the following overview the first number is the number of instances of the annotation and the second number the number of instances of the lingering variant.

- **Nod** (681 & 766 lingering) - The main stroke of the vertical head movement is downwards.
- **Backnod** (428 & 290 lingering) - The main stroke of the vertical head movement is upwards.
- **Double nod** (154 & 4 lingering) - Two repeated head nods of the same intensity.
- **Shake** (17) - Repeated horizontal head movement.
- **Upstroke** (156) - Single vertical movement upwards. This can either occur independently or just before a nod.
- **Downstroke** (43) - Single vertical movement downwards. This can either occur independently or just before a backnod.
- **Tilt** (24 left & 15 right) - Rotation of the head, leaning to the left or right.
- **Turn** (8 left & 11 right) - Turning of the head to left or right direction.
- **Waggle** (7) - Repeated nods accompanied by multiple head tilts.
- **Sidenod** (9 & 2 lingering) - Nod accompanied by a turn to one direction (6 left & 5 right).
- **Backswipe** (18 & 2 lingering) - Backnod which is not only performed with the neck, but also with the body which moves backwards.
- **Sideswipe** (3 left & 5 right) - Sidenod which is not only performed with the neck, but also with the body which moves to that direction.

Keep in mind that head movements are annotated only in areas where a listener response was identified in the first step of the annotation process. Especially turns and tilts occured more often than reflected in these numbers, but the others were not categorized as listener responses.

**EYEBROWS** For the listeners eyebrow raises and frowns were annotated. It was indicated whether the movement concerned one or both eyebrows. When one eyebrow was raised or frowned, it is indicated which eyebrow (left or right) made the movement. In total this layer contains 200 annotations, 131 raises and 69 frowns. These numbers include the annotations in which only one eyebrow was raised or frowning occurred with one eyebrow.

**MOUTH** The movements of the mouth were annotated with the following labels (457 in total): smile (396), lowered mouth corners (31), pressed lips (22) and six other small categories (8). Especially with smiles the end time was hard to determine. If the person was smiling, but increases the intensity of the smile, a new annotation was created.

**SPEECH** For the speakers we collected the results of the automatic speech recognition software SHoUT [81]. For listeners the speech was transcribed. In total 186 utterances were transcribed. The most common utterances were “uh-huh” (76), “okay” (42) and “ja” (29).

**RESPONSES** This annotation layer was created in the third step of the annotation process of the listener. 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). This resulted in 2456 responses. If there was no head movement or vocalization present, either the eyebrow or mouth movement was chosen as onset (whichever came first). The corpus includes 233 responses with mouth movement at the onset and 106 response with eyebrow movement at the onset. In total 2796 responses are in the corpus. Unless otherwise indicated, we have only used the 2456 responses including a head gesture and/or vocalization for the remainder of the thesis.

## 2.3 Consensus Perspective

We set out to record this corpus to get a wider coverage of the response opportunities in an interaction and to be able to analyze the graded optionality of these opportunities. Therefore, we need to combine the three perspectives on appropriate listening behavior into a *consensus perspective*. The consensus perspective can be defined as the complete coverage of all identified response opportunities in a corpus and for each identified response opportunity, the number of listeners that responded to that opportunity.

To create a consensus perspective, we need to identify which responses from different listeners are in response to the same response opportunity. A response opportunity is not a single point in time, but there is a window of opportunity in which the speaker expects the response to be given by the listener. How big this window is, will differ between response opportunities and is mostly controlled by the speaker. The most reliable way to create a consensus perspective is to have annotators group listener responses from different listeners that are in response to the same response opportunity. This was done for 8 out of 32 interactions and a more in-depth analysis

**Algorithm 1** Consensus perspective 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 $firstResponse$
    $thisResponseOpportunity =$ all responses starting in $(tStart + consensus\_window)$
    $lastResponse =$ latest in $thisResponseOpportunity$
    $tEnd =$ start time of $lastResponse$
    $allResponseOpportunities = allResponseOpportunities + [tStart, tEnd]$
    $allResponses = allResponses - thisResponseOpportunity$
  **end while**
  **return** $allResponseOpportunities$

of these interactions will be presented in Chapter 3.

### 2.3.1 Consensus Perspective Algorithm

Since we did not have the time to collect annotations for the whole corpus, we developed an algorithm to create the consensus perspective for all interactions automatically. The algorithm is based on the proximity of listener responses; listener responses that are closely grouped together are considered to be reactions to the same response opportunity. For this we need to specify "closely grouped together" further. We need to define the maximum width of the response opportunity for the algorithm, the so-called *consensus window*. We do not want the algorithm to create response opportunities that included more than one listener response from the same listener. Therefore, we analyzed the recordings and found the minimal gap between two listener responses from the same listener to be 714 ms. To ensure that our algorithm did not group two responses from the same listener, the consensus window was set to 700 ms.

The algorithm is presented in Algorithm 1. 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 start time of the identified response opportunity is the onset of the first response. The end time of the identified response opportunity is the onset of the latest response included in the response opportunity. Note that this means that if a response opportunity with only one response is created, the start and end time of the response opportunity are identical. After a response opportunity is created we continue our forward looking search for the next unassigned response.

In Figure 2.3 an example is given of the consensus perspective building algorithm. At time 1.0 s 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 700 ms. There is a response from listener 2 at time 1.2 s, thus these are

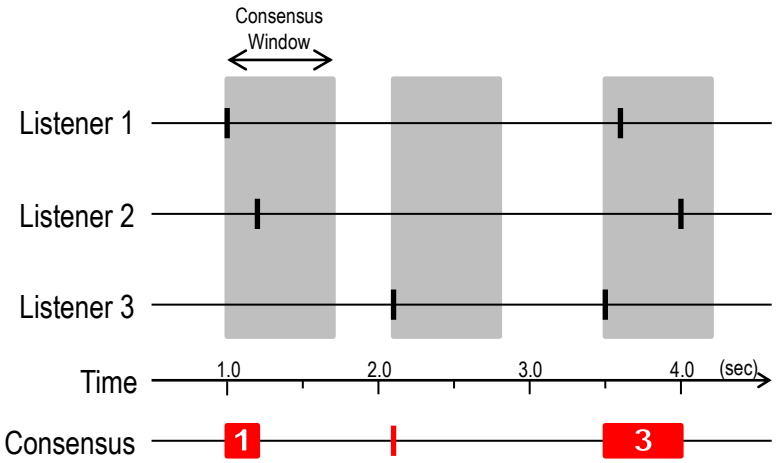


**Figure 2.3:** Example of the consensus perspective building algorithm. The algorithm identifies three response opportunities by grouping the responses from the different listeners that fall within the consensus window. The width of the response opportunity is determined by the start times of the first and last responses within the consensus window.

grouped into consensus instance 1, which starts at time 1.0 s and ends at time 1.2 s. The algorithm continues with the next unassigned response and repeats the process and creates a consensus instance at 2.1 s from the response from listener 3 and from 3.5 s to 4.0 s, by combining the three responses from listener 3, 1 and 2.

The algorithm was applied on all 32 interactions in the MultiLis corpus. From the 2456 responses including a head movement and/or vocalization the algorithm created a consensus perspective identifying 1733 response opportunities. There are 1140 response opportunities with a response from one listener (RO1), 465 response opportunities with responses from two listeners (RO2) and 128 response opportunities with a response from all three listeners (RO3).

### 2.3.2 Coverage of Response Opportunities

One of the reasons we recorded the MultiLis corpus was to increase the coverage of the response opportunities in an interaction. To analyze our success towards this goal, we applied the consensus perspective building algorithm to all subsets of listeners perspectives.

Figure 2.4 illustrates the results of this analysis. If we use only one listener perspectives to identify response opportunities in the interactions we will identify 818 response opportunities on average, depending on which listener perspectives are used. When we use two listener perspective we identify on average 1336 response opportunities, an increase of 63%. Finally using all three listener perspectives a total of 1733 response opportunities are identified, another increase of 30% over the previous number.

So, with the two additional listener perspectives over the one perspective in a traditional corpus, we have more than doubled the coverage of response opportunities in our corpus; from 818 response opportunities to 1733 response opportunities, an increase of 112%.

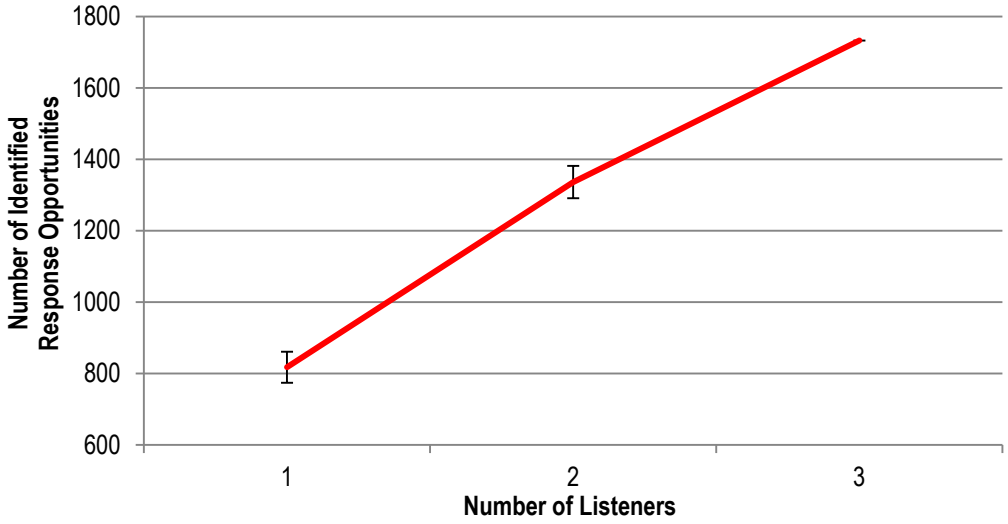


**Figure 2.4:** Graph illustrating the effect of adding multiple listener perspectives on the coverage of response opportunities in an interaction.

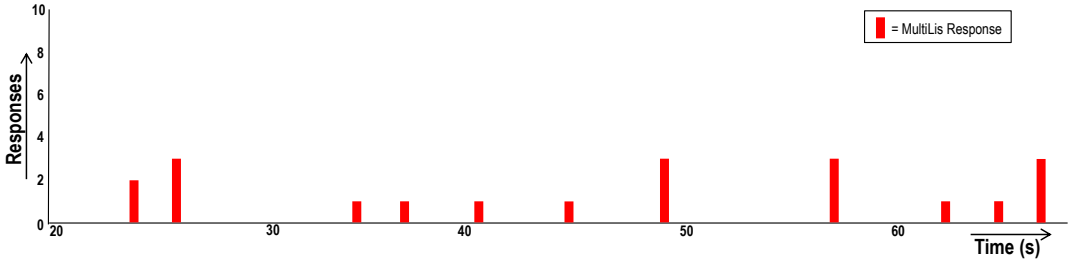


**Figure 2.5:** Sample of the distribution of response opportunities in the MultiLis Corpus.

## 2.4 Conclusion

In this chapter we have presented the MultiLis corpus. For this corpus we recorded interactions between one speaker and three listeners. The three listener were unaware of each other. By combining the listener perspectives of these three listeners into a consensus perspective, we have demonstrated that the coverage of so-called *response opportunities*, moments in an interaction where a listener response is possible or even expected by the speaker, is increased by 112%.

To illustrate the increase in coverage of response opportunities and the graded optionality, we will take a closer look a sample from one of the interactions in our corpus. Figure 2.5 represents a segment of 48 seconds from one of the interactions[4]. It shows the distribution of response opportunities in this segment. The horizontal axis represents time. The response opportunities in these 48 seconds found in the MultiLis corpus are indicated by red bars. The height of these bars represents the amount of recorded listeners that gave a response at this response opportunity.

The segment is taken from an interaction where agreement between listeners is relatively high. In this segment there are four response opportunities with three lis-

[4] 19.1 to 1:06.5 seconds from interaction 9

tener responses, one with two listener responses and six with one listener response. No listener performed a listener response at all these response opportunities. This illustrates that with this corpus we have a more complete view of all the opportunities for a listener response. In the following chapters we will keep returning to this segment.

# 3

# Parasocial Sampling

The previous chapter has shown that recording two additional listeners increases our coverage of response opportunities by 112%. However, we presumably have not yet reached total coverage of all response opportunities. There are still moments that are appropriate moments to respond to as a listener, but none of our listeners have responded at that point in time. More listeners are needed to collect a complete picture of all response opportunities in an interaction and to get a good view of the graded optionality of each response opportunity. However, recording even more listeners in parallel is a complicated and costly operation.

Using recordings of conversations is not the only way in which listener responses have been studied. Watanabe and Yuuki [154] built a voice reaction system, a system that could generate listener responses based on the speaker's voice. To develop their system they used data, where two listeners intentionally nodded to the speaker's speech. The two listeners heard a telephone message uttered by a speaker and they visually nodded in response, which was recorded by using a videorecorder.

Later, Noguchi and Den [114] streamlined this process using keys on a keyboard to record the listener responses instead of acted nods recorded on videotape. They presented several pause-bounded phrases consisting of a single conversational move. Subjects involved in the experiment were asked to hit the space bar of a keyboard if they thought it was appropriate to respond to the stimulus with a listener response. This way, they circumvented the process of having to annotate the videotape.

Finally, Huang *et al*. [79] collected similar data by presenting complete interactions instead of single conversational moves to their subjects. They named this collection method *Parasocial Sampling* (PS), after research into parasocial interaction [77]. This research suggests that individuals can engage in interaction with pre-recorded media as if they were engaged in a natural social interaction.

## 3.1 Parasocial Sampling

In the following section we are going to describe the collection of perspectives using the parasocial method from Huang *et al*. [79]. Parasocial listeners listened to com-

plete interactions from the MultiLis corpus and gave their parasocial perspective on appropriate listening behavior.

The goal of the experiment was two-fold. First and foremost we wanted to collect more perspectives of appropriate listening behavior to increase the coverage of response opportunities in our corpus and be able to gain more insight into the variation of the behavior. Which response opportunities seem to be mandatory, which are preferred by most and which are only occasionally responded to? Discriminating between these different types of response opportunities becomes more reliable with more perspectives. Furthermore, we wanted to compare the parasocial perspectives to the parallel recorded perspectives in terms of numbers of responses given, timing of these responses and the individuality of the captured behavior to assess the validity of the parasocial alternative.

For this experiment we collected parasocial perspectives for 8 out of 32 interactions from the MultiLis corpus. We invited six of the eight original listeners of the data collection and ten additional subjects. The exact procedure for this data collection will be explained in Section 3.1.1. What the impact of these additional perspectives will be on the coverage of response opportunities is presented in Section 3.1.2. Finally these additional parasocial perspectives will then be compared to the original listener perspectives in Section 3.1.3.

### 3.1.1 Procedure

The collection of parasocial perspectives was performed on eight interactions from the MultiLis corpus [1]. Ten months after the original MultiLis experiments we reinvited six of the original listeners in these eight interactions to collect their parasocial perspectives for the same interactions. While watching and listening to the three recordings of the same speakers they listened to earlier, they gave responses through the keyboard. Each time they would give a listener response they were instructed to press the spacebar of the keyboard to collect their parasocial perspective.

Furthermore, we invited ten new participants to collect their parasocial perspectives on these interactions. Each of these participants gave their parasocial perspectives on four interactions. Thus, for each of the eight interactions, we have three original listener perspectives and seven or eight parasocial perspectives. From these perspectives there are five perspectives from the new participants and two or three perspectives from the original listeners, depending on whether one of them was the speaker in that interaction or not.

### 3.1.2 Coverage of Response Opportunities

So, let us take a look at how much the additional parasocial sampling perspectives have increased the coverage of response opportunities. For this we combined the original three listener perspectives with the five parasocial perspectives of the new participants. The parasocial perspectives of the original listeners were not included, because of the unbalanced number of perspectives this created for some interactions and the duplicate nature of these perspectives.

[1] Interactions 9-12, 25-28

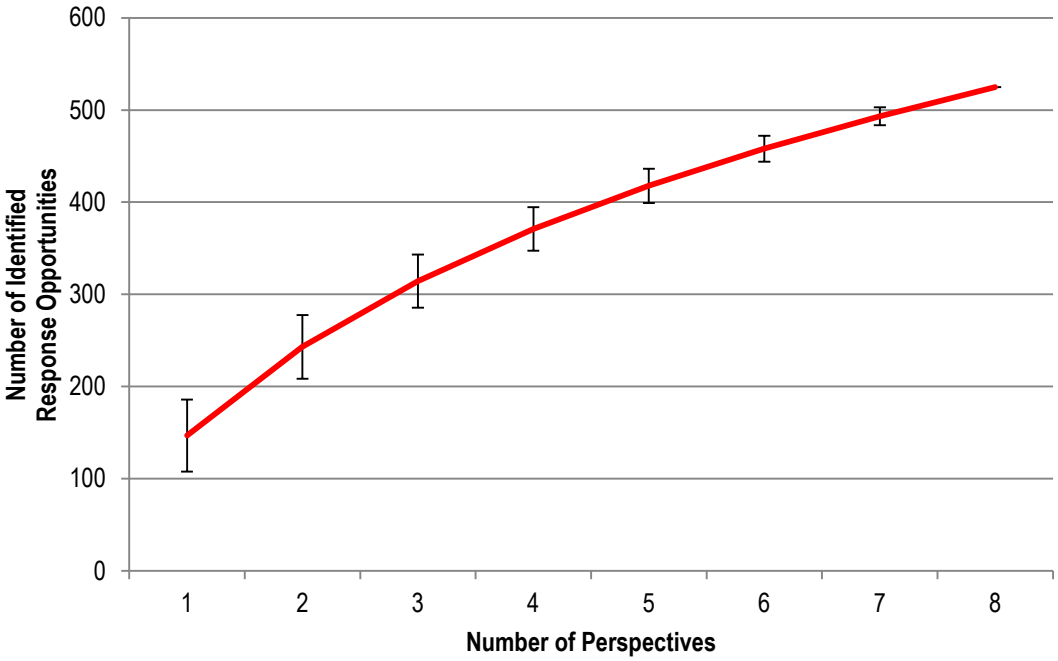


**Figure 3.1:** Graph illustrating the effect of additional parasocial sampling perspectives on the coverage of response opportunities in an interaction.

Initially, we built the consensus perspective using the algorithm from Section 2.3.1. After this one annotator did a manual pass over the created response opportunities and corrected any mistakes the algorithm made according to the annotator's judgment. The mistakes made by the algorithm were always of the kind that the response opportunities created were not inclusive enough. In other words, there were additional responses that belonged to the same response opportunity, which were not included by the algorithm. After this we counted the number of discovered response opportunities for each subset out of the eight perspectives available.

Figure 3.1 shows the increase each additional perspective brings. Similarly to the original three listener perspectives the increase for the first three perspectives for these interactions is around 114%, from 147 response opportunities with one perspective to 314 response opportunities with three. Each additional perspective keeps increasing the coverage, but as can be expected the increase is less with each step. From three to four perspectives the relative increase is still 18%, but for the final step, from seven to eight perspectives, the relative increase is only 6%. Combining all eight perspectives gives us 525 response opportunities, a total increase of 257% compared to only one perspective.

### 3.1.3 Comparison of Parallel Recording and Parasocial Sampling

In this section we will compare the two data collection methodologies; parallel recording and parasocial sampling. We will look at the response rate and timing of each methodology as well as the agreement between the two methodologies.

For the six participants that took part in both the MultiLis experiment and in the

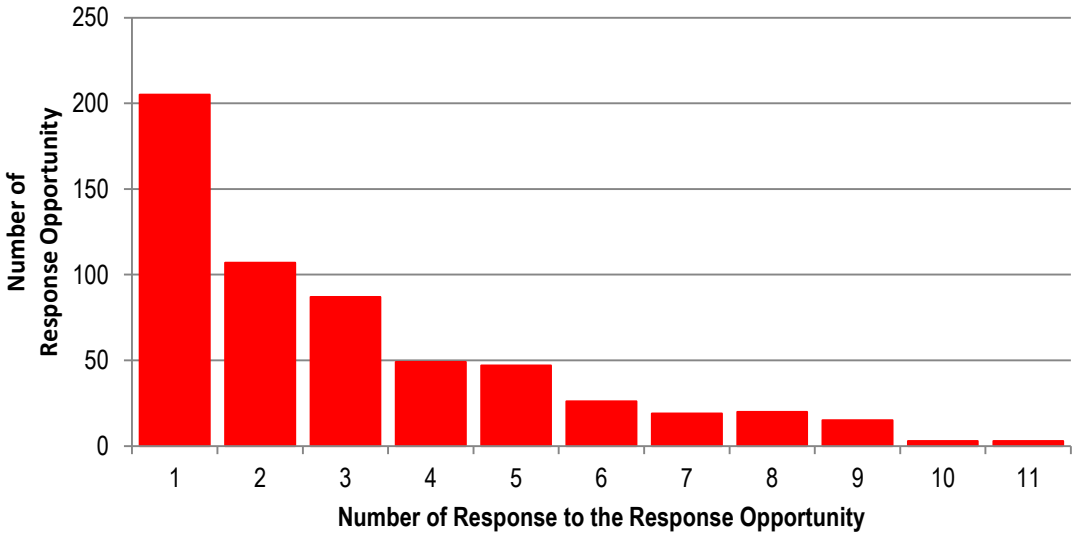


**Figure 3.2:** Histogram of the number of responses to each response opportunity in the consensus perspective of both the MultiLis and parasocial perspectives.

parasocial sampling experiment, their collected perspectives included in both cases on average 5,75 responses per minute. So, in both collection methods the response rate is comparable. The ten participants that only participated in the parasocial sampling collection responded on average 5 times per minute.

For the following analyses we also included the parasocial perspectives of the re-invited partipants in our consensus perspective. This perspective now includes a total of 581 response opportunities. The distribution of the number of responses to each response opportunity in the consensus perspective is presented in Figure 3.2. Most response opportunities have only a few responses, but there are still 15 response opportunities with 9 responses, 3 with 10 responses and 3 with 11 responses. It is interesting to note that all response opportunities in which 10 or 11 people agreed and 4 response opportunities where 9 agreed that a response should be given were from the same interaction (number 10). We will look a bit closer at the reason for this in Section 5.1.

For now we will take a closer look at the timing of the responses to the same response opportunity. What is the window for each response opportunity? Or in other words, what is the range in which the different perspectives respond to the same response opportunity? Furthermore, is there a difference in timing between the responses of the listener perspectives and parasocial perspectives?

In Table 3.1 an overview is presented of the 581 response opportunities identified in the eight interactions by the combination of both methods. The total mean length of a response opportunity is 0.75 seconds and ranges from 0.01 to 2.35 seconds. Note that response opportunities with only 1 response have no duration, since we only have one start time to span the window of opportunity.

Next, we will look at the timing of the responses of the listener perspectives versus those from the parasocial perspective. For this analysis we measured the distance between the start of each individual response and the beginning of the corresponding

| Number of Responses | Number | Mean Duration | Minimum Duration | Maximum Duration |
|---|---|---|---|---|
| 1 | 205 | - | - | - |
| 2 | 107 | 0.35 | 0.01 | 0.96 |
| 3 | 87 | 0.54 | 0.08 | 1.25 |
| 4 | 49 | 0.75 | 0.08 | 1.45 |
| 5 | 47 | 0.93 | 0.22 | 1.69 |
| 6 | 26 | 1.04 | 0.38 | 1.97 |
| 7 | 19 | 1.13 | 0.41 | 1.86 |
| 8 | 20 | 1.22 | 0.58 | 2.18 |
| 9 | 15 | 1.28 | 0.54 | 2.35 |
| 10 | 3 | 1.10 | 1.00 | 1.23 |
| 11 | 3 | 1.40 | 1.08 | 1.86 |
| Total | 581 | 0.75 | 0.01 | 2.35 |

**Table 3.1:** Overview of the response opportunities with minimum, maximum and mean duration for each number of responses.

| | Original Behavior | | |
|---|---|---|---|
| | Listener | Listener | Listener |
| PS | 1 | 2 | 3 |
| Lis 1 | *0.52* | 0.46 | 0.41 |
| Lis 2 | 0.44 | *0.50* | 0.48 |
| Lis 3 | 0.18 | 0.21 | *0.35* |
| PS 1 | 0.42 | 0.42 | 0.41 |
| PS 2 | 0.27 | 0.37 | 0.35 |
| PS 3 | 0.39 | 0.48 | 0.43 |
| PS 4 | 0.24 | 0.26 | 0.52 |
| PS 5 | 0.41 | 0.50 | 0.41 |

| | Original Behavior | | |
|---|---|---|---|
| | Listener | Listener | Listener |
| PS | 5 | 6 | 7 |
| Lis 5 | *0.37* | 0.29 | 0.24 |
| Lis 6 | 0.25 | 0.36 | *0.43* |
| Lis 7 | *0.28* | 0.18 | 0.27 |
| PS 6 | 0.24 | 0.27 | 0.18 |
| PS 7 | 0.18 | 0.26 | 0.39 |
| PS 8 | 0.30 | 0.31 | 0.39 |
| PS 9 | 0.27 | 0.25 | 0.34 |
| PS 10 | 0.25 | 0.17 | 0.16 |

**Table 3.2:** Agreement between the collected parasocial perspectives for each participant and the original listener perspectives (measured in $F_1$ scores). The table is split into two; one for each session of four interactions.

response opportunity. The mean distance of responses from the original listener is 0.28 seconds and for parasocial responses 0.50 seconds. Thus parasocial responses are significantly slower than the original responses (ANOVA: $F = 92, p < 0.01$). So, even though the participant for the parasocial experiment can act as if being involved in the interaction, the combined effect of the conscious nature of the task and unnatural physical action to collect the parasocial sampling, results in a delay of 0.22 seconds.

In our final analysis of this data, we look at how the parasocial perspective of the re-invited listeners compares to their original listener perspective. For this analysis we compared the timing of the response in each of the perspectives to each other. We counted the number of times the responses from one perspective were in response to the same response opportunity as the other perspective. From these counts precision and recall were calculated and combined by taking the weighted harmonic mean into the $F_1$ measure. Agreement varies from 0.18 to 0.52.

The agreement of each comparision is presented in Table 3.2. The table is split into two. The left side of the table is for interactions 9 to 12 and the right part is about interactions 25 to 28. In the first three rows the agreement between the parasocial perspectives of the re-invited listeners is compared to the original listener perspectives. In the bottom five rows the agreement between the new parasocial perspectives and the original listener perspectives is presented.

If we want to know with which original listener perspective a parasocial perspective has the most overlap, we need to look for the highest number on each row. In the table the highest number on each row is in italics. For the parasocial perspectives of the re-invited listeners this is in four out of six cases their own listener perspective with $F_1$ scores ranging from 0.27 to 0.52. Thus, for four out of six listeners the closest match with their parasocial behavior is their own recorded behavior.

If we want to know with which parasocial perspective the recorded listener perspective has the most overlap, we need to look for the highest number in each column. In the table the highest number in each column is underlined. Again in four out of six cases the closest match with their recorded listener perspective is their own parasocial perspective, with $F_1$ scores ranging from 0.36 to 0.52.

The fact that the parasocial perspective of a person collected ten months after the original recordings is the closest match among all collected parasocial perspectives in most of the cases, suggests that each listener has an individual preference to which response opportunity they respond. Furthermore, it proves that collecting listening behavior through the parasocial sampling method is capable of capturing this individual preference. This makes it a viable substitute for actual recording of listening behavior with regards to timing of listener responses. A downside of the method is the lack of visualization of the behavior, so no analyses can be performed on how this listener would give the listener response (vocal/visual/both, upwards/downwards nod, amplitude of the behavior, etc.).

### 3.1.4 Conclusion

In this section we have presented the collection of parasocial perspectives. These parasocial perspectives were collected by asking people to give listener responses as if they were interacting with the presented speaker, through means of the space bar on a keyboard. We have demonstrated that these parasocial perspectives closely match recorded listener responses in terms of number of responses given and the ability to capture the individual preference for listening behavior of the person. The responses were given with a small delay of 0.22 seconds on average. We have shown that collecting in total 8 (listener and/or parasocial) perspectives increased the coverage of response opportunities up to 257% increase over one recorded listener.

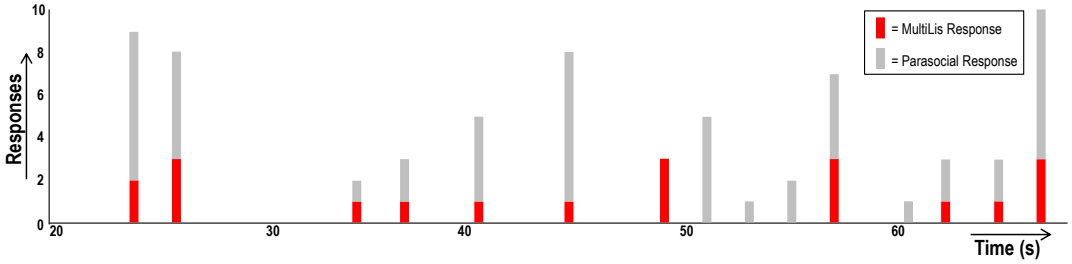


**Figure 3.3:** Sample of the distribution of responses opportunities in the MultiLis Corpus and parasocial responses.

To illustrate the impact of the extra parasocial perspectives on the coverage of response opportunities in the MultiLis corpus in Figure 3.3 we will take a look at the same 48 seconds as in the previous chapter. In gray the responses from the collected parasocial perspectives are added to the responses from the MultiLis corpus. The participants provided a parasocial response to almost all the response opportunities found in the previous study with the exception of the response opportunity just before 50 seconds. Interestingly this response opportunity was responded to by each listener in the MultiLis corpus. Furthermore there are four new response opportunities of which one was responded to by five participants. This demonstrates both the validity of the parasocial collection method, the parasocial responses correlate with the listener responses from the corpus, as well as the increased coverage we have gained by collecting these parasocial responses.

## 3.2 Individual Perceptual Evaluation

So far we have collected data to get an as complete as possible picture of the response opportunities in an interaction by combining multiple perspectives on what appropriate timing for listener responses are. In the eight interactions from the previous study we have found 581 response opportunities. These response opportunities cover 435 seconds of the 2054 seconds these interactions last in total. So, in these interactions 21% of the time there is a response opportunity according to the perspectives we have collected. So does this automatically mean that the remaining 79% of the time it is an inappropriate time to give a listener response? Is a listener response generated at such a time perceived as wrong? Does a listener response given at such a time disrupt the conversation?

To answer these questions we use the Individual Perceptual Evaluation method. In this method we generate virtual listening behavior in reaction to a recorded speaker and let participants judge each generated listener response on appropriateness. During watching the interaction between recorded speaker and generated listener the participants are instructed to press the space bar when they judge the generated response to be inappropriate. We specifically instructed the participants to make a judgment based on timing (see for more details on the procedure Section 3.2.1). We thus collected a parasocial perspective of the inappropriate timing of listener responses. However, due to the setup the perspective is discrete and not continuous. Only the

moments where a listener response was generated are judged and no judgment is available for the moments in-between.

The listener responses of the embodied conversational agent in these interactions were mostly generated at known response opportunities and sometimes generated at moments where there was no response opportunity according to the consensus perspective. This way the listening behavior of the embodied conversational agent still appears natural and the participant is only presented a potentially wrong listener response sometimes. More details on the stimuli will be presented in Section 3.2.2.

We expect the participants to judge most listener responses generated at moments where there was no response opportunity according to the consensus perspective as inappropriate. Furthermore, we expect most listener responses generated at response opportunities to be judged as appropriate, but not necessarily all moments. The results for this analysis will be presented in Section 3.2.3.

### 3.2.1 Procedure

We invited eight subjects to participate in the study. After a short introduction on what listening behavior is and what the function of the behavior is, they watched the eight interactions between a recorded speaker and a generated virtual listener. They were instructed to pay attention to the timing of the listener responses and judge each listener response on whether or not they thought the listener response was appropriately timed. When a listener response was inappropriate according to their judgment they were instructed to press the space bar on a keyboard (a *yuck response*). The participant had the option to replay the video, which would result in a loss of all collected judgments for that video so far.

### 3.2.2 Stimuli

We presented subjects with clips of a speaker from the MultiLis corpus in interaction with a virtual listener, animated using the BML realizer Elckerlyc [142]. We used the same eight interactions as in the previous study. As listener responses the virtual listener performs only head nods and every time the same head nod. The timing of the head nods is based on the consensus perspective from the previous studies.

Two-thirds of the head nods (182) were generated at known response opportunities and one-third of the head nods (90) were generated between these known opportunities. The 182 head nods generated at known response opportunities (or *at-head-nods*) were placed at response opportunities with at least four response from the multiple perspectives. The 90 head nods generated between known response opportunities (or *between-head-nods*) were placed in the 90 biggest gaps between the *at-head-nods*. Within these biggest gaps they were placed in the biggest gap between all response opportunities in the consensus perspective. This includes all response opportunities in the consensus perspectives, including the response opportunities with less than four responses.

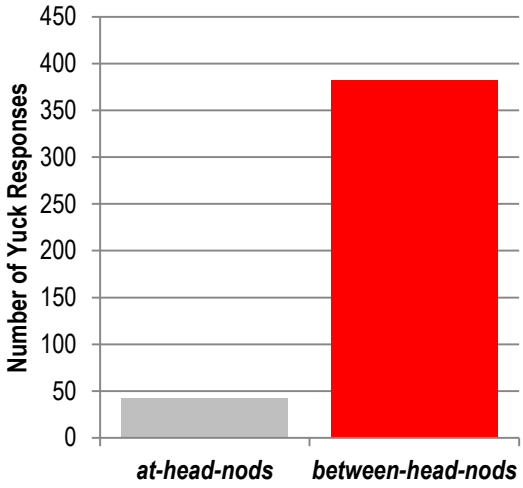


**Figure 3.4:** The number of yuck responses in reaction to the head nods generated at different times. The head nods generated at response opportunities were perceived as more appropriately timed than head nods generated between response opportunities.

### 3.2.3 Results

We analyzed the number of yuck responses, judgements that a generated listener response is inappropriately timed, in reaction to the two different types of generated head nods, the *at-head-nods* and *between-head-nods*. On average each participant judged 53 out of 272 head nods as inappropriate, for a total of 424 yuck responses. As expected the participants judged the *between-head-nods* as less appropriate than *at-head-nods* (see Figure 3.4). Forty two yuck responses were in reaction to *at-head-nods* and 382 were in reaction to *between-head-nods*. The 42 yuck response in reaction to *at-head-nods* were in reaction to 29 individual *at-head-nods*. So, 16% of the *at-head-nods* received a yuck response from at least one participant. Four of these *at-head-nods* were yucked three times, five were yucked twice and the other 22 were yucked once.

For each of the generated *between-head-nods* we counted the number of yuck responses. Figure 3.5 shows the histograms of the 382 yuck responses in reaction to these 90 *between-head-nods*. Most of the *between-head-nods* (56 out of 90) were yucked by at least half of the participants. There were three *between-head-nods* which were yucked by each participant. There were eight *between-head-nods* which were found appropriate by each participant (9%), even though the combination of multiple perspectives in the previous studies had not detected a response opportunity at that time.

## 3.3 Conclusion

In this section we have presented the individual perceptual evaluation method to collect perspectives on inappropriate timing of listener responses. Participants evaluated in an interaction between a recorded speaker and a generated listener head nods generated at known response opportunities and between response opportunities on their timing. We have shown that 16% of the head nods generated at known response

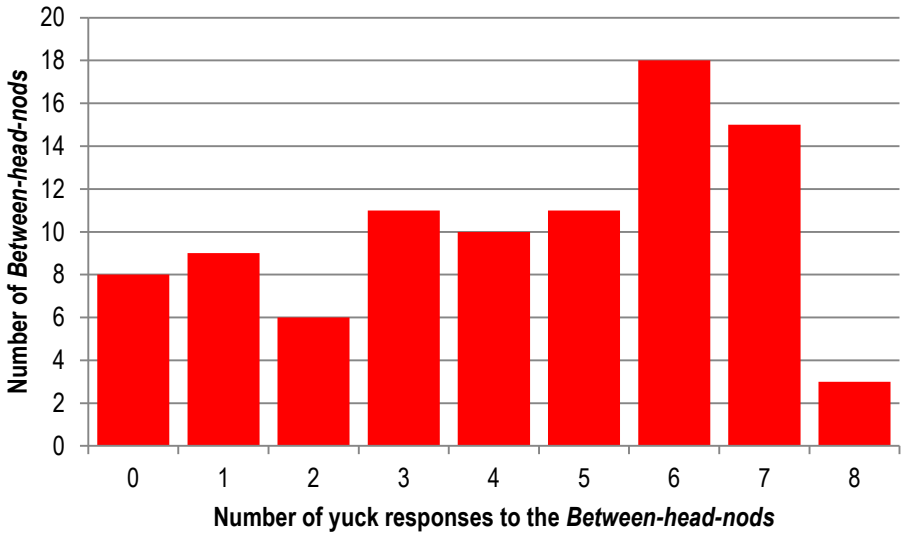


**Figure 3.5:** Histogram of the number of yuck responses to each *between-head-nod*.

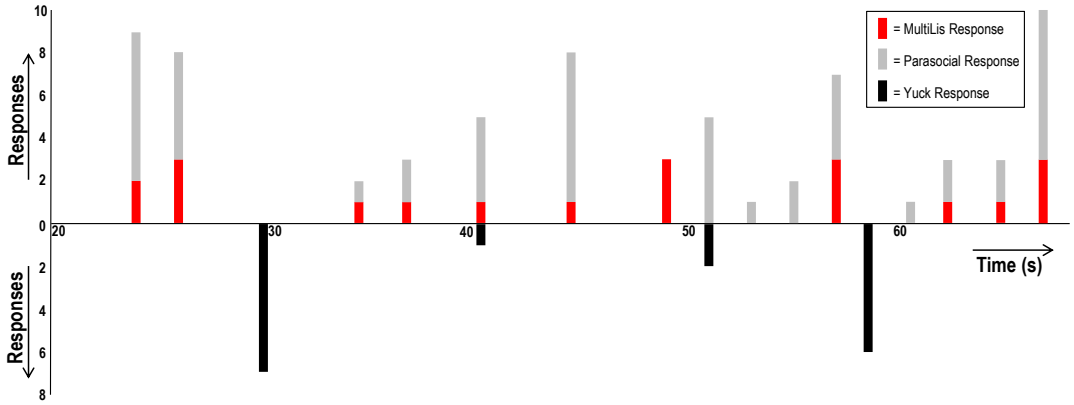


**Figure 3.6:** Sample of the distribution of responses in the MultiLis corpus, parasocial responses and the yuck responses.

opportunities were still judged as inappropriately timed, while 9% of the head nods generated between response opportunities were judged as appropriately timed.

How the collected perspectives on inappropriate timing has affected the segment of 48 seconds we looked at before, is presented in Figure 3.6. In this figure we have added the yuck responses below the previous responses as negative responses, represented as black bars. Note that only a head nod was generated and evaluated at response opportunities with at least four MultiLis or parasocial responses. The moments just before 30 seconds and just before 60 seconds were the only generated head nods between response opportunities in this segment. So, in this segment there were no head nods generated between response opportunities that nobody judged as inappropriate. There was a moment that was found to be a response opportunity by the parasocial perspectives, but was judged as inappropriate in this evaluation. We will take a closer look at this segment at a later stage when we analyze what goes on in the interaction at these moments.

# Part II

# Analysis of Listening Behavior

# 4

# Listener Equality

As mentioned before communication is a collaborative process. Both speaker and listener coordinate their actions and behavior to transfer information to each other. While the speaker shares information, the listener assesses this information and signals whether the information sent by the speaker has been received, understood and/evaluated. This way a common ground of knowledge that both participants share is established.

Research into the role of the listener in this process has a long tradition. In an experiment where subjects were required to reproduce various geometric figures on the basis of descriptions from the experimenter, Leavitt and Mueller [97] showed that increasing feedback from the listener resulted in increasing accuracy of the reproduction. Thus, through the response from the listener the speaker is able to tailor the details unclear to the listener and provide more accurate descriptions. In a repeated version of a similar experiment, Krauss and Weinheimer [92] found that the speaker used more words to describe the graphic designs in successive trials, when the listener gave no feedback to the listener in between trials. Krauss *et al*. [91] showed that delaying feedback in the same experiment also resulted in the speaker using more words to describe the graphic designs. An explanation for these results lies in the fact that the speaker receives no or delayed feedback from the listener, so the speaker begins to doubt his own explanation and gives additional details or rephrases.

Kraut *et al*. [93] conducted a study where two listeners were listening to the same summarization of a movie by a speaker. One of the listeners could provide feedback to the speaker, while for the other listener the audio channel back to the speaker was disconnected without him knowing. They demonstrated that the listener that could provide feedback had a better understanding of the summarized movie than the disconnected listener. Bavelas *et al*. [7] let subjects tell a close-call story to a listener who either was distracted from the narrative content while listening or not. When the listener was distracted fewer responses were given and the speakers told their stories significantly less well.

Gratch *et al*. [61, 62] reproduced similar results in human-agent interactions. They researched the influence of a responsive virtual listener as opposed to an unre-

| | | Seen by speaker? | | |
|---|---|---|---|---|
| **Noticed illusion?** | | Correct | No Idea | Wrong |
| Yes | 14 | 6 | 3 | 5 |
| No | 17 | 5 | 1 | 11 |
| All | 31 | 11 | 4 | 16 |

**Table 4.1:** The table shows the reported answers to questions in the post-questionnaire regarding the manipulation. 14 out of 31 participants reported that they noticed that they were not always seen. In total 11 out of 31 participants could correctly identify the interaction in which they were seen by the speaker. This is hardly above chance (= 33%)

sponsive virtual listener. Human speakers interacting with a responsive virtual listener used fewer words, had fewer pause fillers and fewer disfluencies. The speakers also felt a greater sense of rapport with the responsive virtual listener.

All these experiments show the role of the listener as vital to the success of the interaction. Feedback from the listener results in more effective communication showing in the number of words needed to transfer information and level of understanding of the listener giving the feedback.

The setup of our MultiLis experiment was very similar to the discussed experiment of Kraut *et al*. [93], but now in a multimodal setting. The listeners could not only hear the speaker, but also see them. The speaker, however, could only see one of the listeners. So, there was two-way communication between the speaker and one of the listeners, but only communication *from* the speaker *to* the two other listeners.

In the previous chapter we combined the three listener perspectives as if they were equal. However, given the findings reported in the discussed literature this may not be valid. The fact that the speaker could not acknowledge the listener responses of two listeners or shape the presentation to their needs, may have had an impact on the behavior or task performance of the listener. In this chapter we will analyze what effect the closing of the interaction loop for the two concealed listeners had.

We will analyze whether they noticed the closing of the interaction loop (Section 4.1), whether the closing of the interaction loop affected their task performance (Section 4.2) and/or their behavior (Section 4.3). We will summarize our findings in Section 4.4.

## 4.1 Self Report

We start by analyzing the perception of the listeners of the interactions. Did the listeners notice that the speaker could not see them in two out of three listener interactions? Did they report a higher rapport between them and the speaker when they could be seen?

We told our participants that they would have four one-on-one interactions, while in fact they were having four one-on-three interactions. Out of those four interactions they were once speaker, once the displayed listener and twice the concealed listener. During the post-questionnaire after the experiment we explained the setup to the participants and asked them if they had noticed this illusion. Table 4.1 shows the results

of this manipulation check. Out of 31 participants 14 reported that they had noticed that they were not always seen (1 participant failed to fill out the post-questionnaire). Out of those 14 only six could correctly identify in which conversation they were seen by the speaker, five guessed a wrong interaction and three reported to have no idea. So, even though they noticed, only about half of them were able to identify the correct interaction. Of the 17 participants that did not notice, five participants guessed the correct interaction in which they were the displayed listener, 11 made the wrong guess and one reported to have no idea. So, in total 11 out of 31 participants correctly identified this conversation. This is hardly above chance (= 33%).

After each interaction the listeners filled out a questionnaire measuring the rapport between the listener and the speaker. The measure consisted of 10 5-point Likert scale questions and was an adapted version of the rapport measure used by Gratch *et al.* [61] with additional questions from the Inventory of Conversation Satisfaction scale from White [157]. Some sample questions are "I was able to motivate the speaker to tell his story well" and "The speaker paid attention to me".

The displayed listeners reported a significantly higher rapport rating for the interactions than the concealed listeners (3.39 vs 3.05 respectively, $p = 0.014$). This is in line with what one would expect, as the speaker did not respond to concealed listeners.

## 4.2 Task Performance

During the four interactions the listeners needed to remember as many details as possible from the summarizations or recitations the speaker gave in the interaction. Kraut *et al.* [93] found that listeners that could give feedback to the speaker remembered more details about a summarized movie than a listener who heard the same summarization, but could not give feedback. In this section we will analyze whether we have reproduced these results in our experiment.

In this experiment listeners either needed to be able to answer 16 multiple choice questions or to reproduce a recipe as accurately as possible. We measured the task performance on these tasks and ranked the three listeners among themselves. More details on this can be found in Section 2.1.3.

To see whether the results in our experiment are in line with these results, we analyzed the rankings of the listeners. We used ANOVA to test for significance of results. The results are presented in Figure 4.1. The mean ranking for displayed listeners is in gray and for concealed listeners in red. Results are presented for all tasks combined, for the video task only and for the recipe task only.

The mean ranking is better for displayed listeners (1.66) than for concealed listeners (1.98), but not quite significant, $F(1, 94) = 3.59, p = 0.06$. However, if we take a look at the two different tasks individually we see that for the video task there is a marginally significant difference between the mean ranking of the displayed listeners (1.50) and the concealed listeners (1.97), $F(1, 46) = 4.00, p = 0.05$. For the recipe task the mean ranking of the displayed listeners (1.81) is a little better than the concealed listeners (2.00), but the difference is not significant, $F(1, 46) = 0.53, p = 0.47$.

Overall, we can say that our data corroborates the findings of Kraut *et al.* [93]

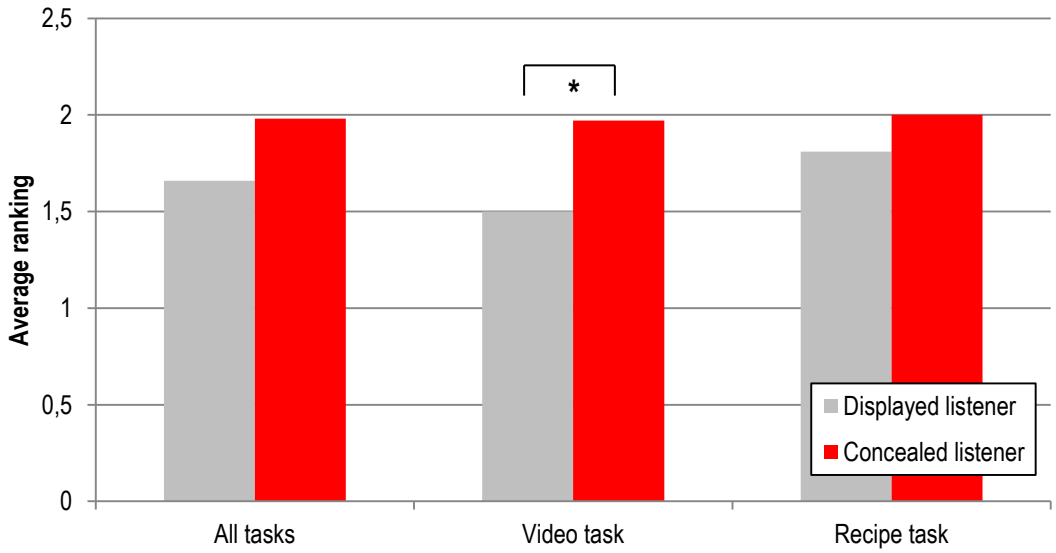


**Figure 4.1:** The average ranking in terms of task performance of displayed (gray) and concealed (red) listeners for all tasks (left), video summarization task (middle) and recipe task (right). Only the difference for the video task is statistically significant ($* : p < 0.05$).

even though they are not as convincing as the results in their paper. However, their setup of the task and measurements were more sophisticated than ours. The speaker had to summarize a whole movie instead of a seven minute short. Furthermore, in their experiment some of the listeners were shown some scenes from the movie and a very similar movie to prime them with confusing information.

## 4.3 Behavior

So far we have seen that some people noticed that they were not seen by the speaker, but had a hard time identifying the interaction in which they were seen by the speaker. The displayed listeners also reported a better rapport with the speaker. Furthermore, we have seen that the displayed listeners had a slightly better performance in the task.

But has the setup of the experiment also changed the behavior of the concealed listeners opposed to the displayed listeners? To answer this we will analyze the behavior of the listeners. We will analyze the behavior based on the annotations made on the MultiLis corpus in Section 4.3.1. We also did a perception study to check whether observers could identify the displayed listener among the three. This study will be presented in Section 4.3.2.

### 4.3.1 Behavior Analysis

In this section we will analyze the behavior of the displayed listeners compared to the behavior of concealed listeners. Specifically we will look at the number of responses given. For these analyses we looked at the 2796 responses used earlier.

Looking at the number of responses per minute, the displayed listeners gave 7.7 responses per minute and the concealed listeners gave 6.8 responses per minute.

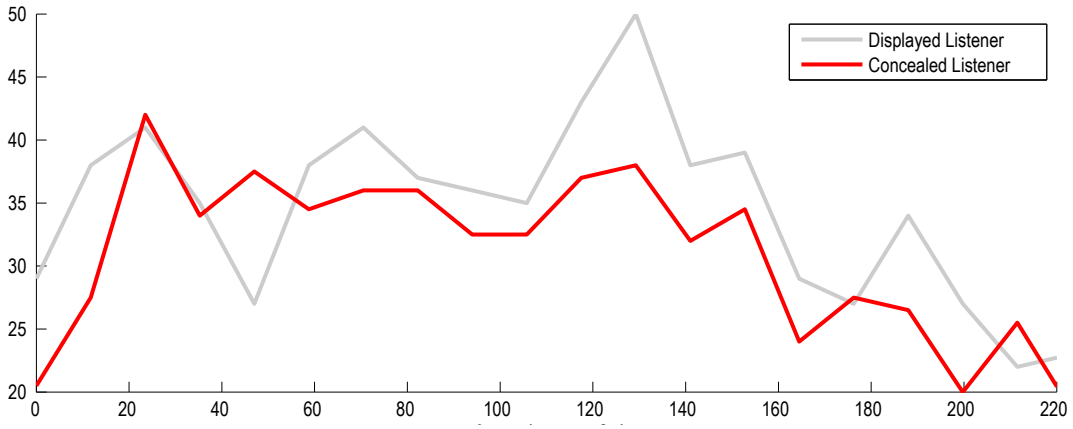


**Figure 4.2:** Graph displaying the distribution of listener responses for both displayed and concealed listeners. It illustrates that although the displayed listeners (gray) usually give more responses than concealed listeners (red), but the difference does not increase over time, if concealed listeners start realizing they are not being seen by the speaker.

This is on average 12% fewer responses from the concealed listeners compared to displayed listeners. However, this difference is not statistically significant, due to the large variance between participants ($p = 0.33$).

If the concealed listeners were to (consciously or unconsciously) notice that they were not seen by the speaker, they would have done this as the interaction progressed. At the beginning they would not know that they were concealed; they may only have noticed that the speaker did not react to their response later on and have given fewer responses as a result of that. If this was true, we expected the difference in response rate between displayed listeners and concealed listeners to increase over time.

To test whether this was the case or not, we have plotted the number of responses over time for the displayed listeners (continuous) and the concealed listeners (dotted) in Figure 4.2. We only used the 15 interactions which lasted longer than 4 minutes. We divided the first 4 minutes into 20 windows and counted how many responses the listeners gave within that time frame. It shows that the number of responses from the concealed listeners is usually smaller than from the displayed listeners, but we do not see that the gap between the two lines increases over time.

### 4.3.2 Perception Study

So, the objective data has not given us significant results which discriminate the behavior of the displayed listener from the concealed listeners. There are indications that concealed listeners give fewer responses, but these results are not significant due to the large variance between listeners. Possibly the changes in behavior are more subtle than one can detect by analyzing objective data. Humans are very capable of noticing subtle changes in behavior. It may only be one precisely timed head nod which discriminates the displayed listener from concealed listeners. This one head nod will get lost in the numbers of an objective analysis, but humans are highly susceptible to such nonverbal cues. Therefore, we performed a perceptive study where we asked observers to look at the interaction between the speaker and the three lis-

teners, with the task to point out the displayed listener.

We invited 16 participants, recruited at the faculty, to participate in the study. The corpus was split into 48 segments ranging from 1:45 to 3:40 minutes. Each participant was shown, through a webpage, 6 segments of the corpus. In the segments the speaker was presented in the top left corner. The three listeners were positioned in the other three corners of the screen. We varied the corner in which the displayed listener was placed. They could pause and repeat the whole or part of the segment. For each of the segments they were asked to identify the displayed listener among the three listeners. Every segment was presented to two participants.

In total the participants answered correctly in 43 of 96 segments (45%), which is significantly better than chance ($P(X >= 43) = 0.01$, with an a priori chance of 33%). The numher of correct answers varied from 0 to 6. There were 8 segments where both participants identified the correct listener, 27 where one participant identified the correct listener and 13 where none of the participants identified the correct listener. Informal interviews with the four participants who identified the correct displayed listener at least 4 out of 6 times revealed that their strategy was to look for reactions of the speaker to one of the listeners. The listener to which the speaker reacted was chosen to be the displayed listener. They especially paid attention to the timing of smiles. They looked for moments where a speaker reacted to a smile by one of the listeners (by smiling or any other type of reaction). The listener who the speaker reacted to must have been the displayed listener.

## 4.4 Conclusion

In this chapter we have analyzed what impact the closing of the interaction loop has had on the two concealed listeners opposed to the displayed listeners. With regards to their own perception about 30% of the participants claimed to have noticed the closing of the interaction loop after being informed, but only half of them could identify the correct interaction in which they were a displayed listener. The participants reported a higher rapport with the speaker when they were the displayed listener than when they were the concealed listener. Furthermore, we have seen that the displayed listeners had a slightly better performance in the task.

However, the change in behavior is minimal. Even if the listeners noticed that they were in the concealed listener condition, they still behaved naturally enough that observers had a hard time identifying who was the displayed listener among the three listeners. Only some observers succeeded in identifying the displayed listeners. The observers did this by paying close attention to the interaction of behaviors between speakers and listeners. In terms of number of listener responses given the difference was not significant. However, there was a trend that concealed listeners gave fewer responses.

Overall, we feel that we have succeeded in capturing natural behavior from the concealed listeners which we can regard as equal to the behavior of displayed listeners for the forthcoming conversational analyses.

# 5

# Conversational Analysis

In the previous chapters the similarity between listeners was analyzed. This was done by first creating a consensus perspective of the collected multiple perspectives. This was followed by an analysis of various other aspects such as the rapport felt, task performance and behavior. Even though similarities were found, no two listeners were found to be identical. From the perspective of response opportunities only a minority of the response opportunities was responded to by all of the listeners. The number of responses each listener gave during the three interactions they were involved in varied between 2.5 and 15.9 responses per minute (mean was 7.5 responses per minute).

Each listener was involved in three interactions. When the number of responses the listener provided was solely caused by the choices and preferences of the listener, independently of the speaker, on average the listener provided the same number of responses in each of these three interactions. The height of the bars in the right graph in Figure 5.1 shows that this was not the case. The interactions with the least number of responses from each listener, the interactions with the most responses and the ones in-between were grouped and the mean responses per minute for each group were calculated. The right graph in Figure 5.1 shows that for the group the number of interactions with the least responses is 3.9 responses per minute, for the middle group 6.9 responses and the group with the most response, 10.5 responses per minute. So, there was a significant variation within the behavior of the listener, caused by the speaker ($p < 0.001$ on a paired-sample *t*-test between all groups).

The speaker is not the sole factor here either. The height of the bars in the left graph in Figure 5.1 shows that there is a variation between the three listeners within a session. Calculating the mean responses per minute for each group of listeners from each session with the least responses per minute (= 4.8), most responses (= 9.4) and the group in-between (= 7.0), we get the left graph in Figure 5.1. The variation is not as big, but still significant ($p < 0.001$ on a paired-sample *t*-test between all groups). This suggests that the combination of speaker and listener determined the number of responses the listener gave, where the speaker had a little more influence on this.

So, this leaves us with the question what did the speaker do to influence the number of responses the listener gives. This is what we will analyze in this chapter through means of conversational analysis. The analysis of the recorded interactions in

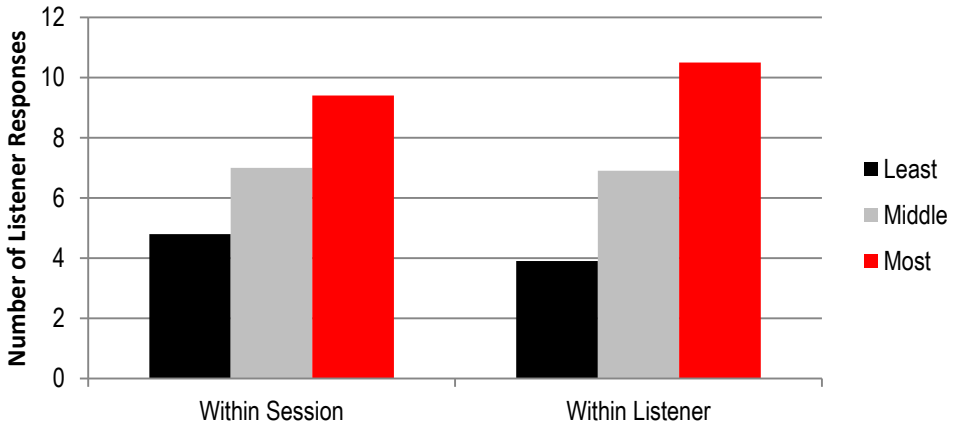


**Figure 5.1:** The left figure shows the average number of responses per minute grouping together the three listener *within a session* into a group of listeners with the least responses, the listeners with the most responses and the listeners in-between. This illustrates that within the three listeners the number of responses varies quite a lot. The right figure shows the average number of responses per minute when looking at the three sessions of *each individual listener* and when grouping them into the sessions with the least responses, the listeners with the most responses and the listener in-between. This illustrates that the response behavior of each listener is highly influenced by the actions of the speaker.

the MultiLis corpus aims to find patterns in the behavior of the speaker in the seconds before the listener gives a response. This will be done in order to find out what cues a speaker gives to make certain response opportunities more compelling to respond to than others. Furthermore, an analysis will be made whether listeners agree on the form of the listener response when they reacted to the same response opportunity.

The chapter will start with an analysis of what the speaker says in the preceeding utterance in Section 5.1. Then the relation between response opportunities and pauses will be looked at in Section 5.2 and the energy of the speech signal in Section 5.3. In Section 5.4 at the preceeding pitch contours will be subject to investigation. This is followed by an analysis of the relation with eye gaze in Section 5.5. Finally, the form of the head gestures the listeners showed and any differences or agreements in that aspect will be analyzed in Section 5.6.

## 5.1 Content

The previous studies collected multiple perspectives on response opportunities, perspectives on both appropriate (Chapter 2 and Section 3.1) and inappropriate (Section 3.2) moments for listener responses. In the combined consensus perspective three type of moments can be identified. These types are *high agreement* (positive or negative), *controversial* and *neutral* moments. The *high agreement* moments have either positive *or* negative responses, the *controversial* moments have positive *and* negative responses and *neutral* moments have neither positive *nor* negative responses. The analyis of these type of moments will be done by presenting several transcriptions of

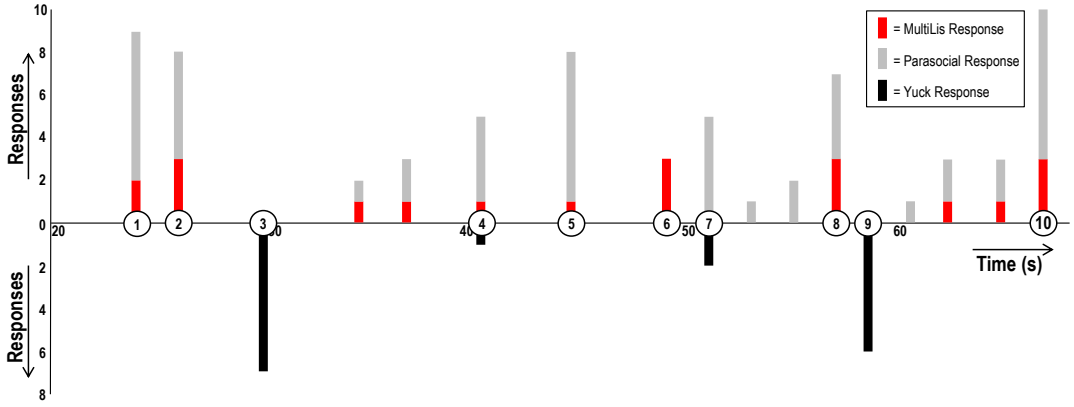


**Figure 5.2:** Sample of the distribution of responses in the MultiLis corpus, parasocial responses and the yuck responses.

these moments and discussing the timing of the responses in relation to the context.

First are the response opportunities with *high agreement*; moments where most perspectives agree these are appropriate or inappropriate moments to provide a listener response. For this we analyze the segment in Figure 5.2 and see what actually happens in the interaction. This segment is taken from an interaction where the speaker recites a recipe for risotto with mushrooms. In this segment the speaker is halfway through the ingredient list. The transcript is presented in Table 5.1. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 5.2. The *high agreement* moments in this segment are 1, 2, 5, 8 and 10 (positive), and 3 and 9 (negative).

The response opportunities 1 and 10 are not in reaction to a summarizing statement. Both statements summarize the previous ingredients with a mnemonic device to help them memorize the ingredients by summarizing the numbers mentioned (1) or by adding up the weights to a round figure (10). Beside the verbal cues, the speaker also makes iconic gestures to accompany the summarizing statements.

The other three *high agreement* response opportunities in this segment (2, 5 and 8) are all in reaction to a refining statement in which a previously mentioned ingredient is more precisely described: the oil is specified as being olive oil (2), the amount of thyme is specified (5) and the precise weight of the rice (8). The other ingredients (like the garlic and onion) are also acknowledged with a listener response by some, but agreement between individuals is much lower in these cases (see the unnumbered response opportunities in Figure 3.6).

The moments with *high agreement* in negative yuck responses (3 and 9) are both mid-sentence. They are not placed near or after the end of a grammatical clause, which was identified as a cue by Dittman and Llewellyn [40], but instead are placed during or directly after the theme of the sentence. So, no new information has been mentioned by the speaker yet (rheme) and the listener response is premature. Furthermore, moments with *high agreement* in negative yuck responses are moments after long silences of at least 2 seconds, moments in between the article and the noun, and moments shortly (within 1.5 seconds) following another listener response.

The moments 6 and 7 are an interesting case that illustrates the influence the delay

**Table 5.1:** Transcript of the segment displayed in Figure 3.6. The numbers in the rightmost column correspond to the response opportunities with the same number in Figure 3.6.

| | | | | |
|---|---|---|---|---|
| 19.1 - 20.8 | twee eetlepels | two tablespoons | | |
| 20.9 - 22.3 | olie. Dus één liter | oil. So one liter | | |
| 22.5 - 23.3 | twee en twee | two and two | | |
| | | | 24.1 | **1** |
| 24.3 - 25.2 | olijfolie | olive oil | | |
| 25.3 - 25.8 | natuurlijk | of course | | |
| | | | 25.9 | **2** |
| 27.9 - 28.7 | uhm | uhm | | |
| 29.6 - 30.1 | je hebt | you've got | 29.9 | **3** |
| 30.3 - 32.9 | verder voor de seasoning | furthermore for the seasoning | | |
| 33.4 - 34.5 | één teentje knoflook | one clove of garlic | | |
| 35.6 - 36.4 | één ui | one onion | | |
| 37.7 - 40.1 | uh twee stengels bleekselderij | uh two sticks of celery | | |
| | | | 40.1 | **4** |
| 42.0 - 42.8 | uh tijm | uh thyme | | |
| 42.9 - 43.9 | één handjevol tijm | one handful of thyme | | |
| | | | 44.4 | **5** |
| 46.4 - 49.0 | en natuurlijk uh heel veel paddestoelen | and of course a lot of mushrooms | 49.0 | **6** |
| 49.1 - 50.1 | 500 gram | 500 grams | | |
| 51.0 - 51.6 | en | and | 51.0 | **7** |
| 51.9 - 52.8 | uhm | uhm | | |
| 53.2 - 54.4 | natuurlijk de rijst | of course the rice | | |
| 55.2 - 57.0 | 400 gram rijst | 400 grams rice | | |
| | | | 57.3 | **8** |
| 57.8 - 58.0 | dus je hebt | so you've got | | |
| 58.4 - 61.9 | uh 500 gram paddestoelen | uh 500 grams mushrooms | 58.6 | **9** |
| | 400 gram rijst | 400 grams rice | | |
| 62.4 - 65.3 | en 100 gram parmezaanse kaas | and 100 grams parmesan cheese | | |
| | dus in totaal | so in total | | |
| 65.4 - 65.7 | mooi | nicely | | |
| 66.0 - 66.5 | één kilo | one kilo | 66.5 | **10** |

**Table 5.2:** Transcript of the most controversial response opportunity in the collected data, with 6 positive responses (3 MultiLis and 3 parasocial) and 3 negative yuck responses.

| | | | | |
|---|---|---|---|---|
| 29.0 - 31.1 | het moment dat hij boven komt, uhm | the moment he arrives at the top, uhm | | |
| 31.6 - 32.1 | oh wacht | oh wait | | |
| 32.3 - 32.9 | helemaal verkeerd | that's wrong | | |
| | | | 33.3 | **11** |

**Table 5.3:** Transcript of a neutral response opportunity where no positive and no negative responses were recorded.

| | | | | |
|---|---|---|---|---|
| 30.5 - 34.1 | uh, volgende list moet ie verzinnen | uh, he has to come up with a new trick | | |
| | hij gaat vanaf | he goes from | | |
| 34.6 - 35.4 | uh | uh | | |
| 35.5 - 36.1 | een tegenoverliggend gebouw | an opposing building | 36.1 | **12** |
| 36.1 - 40.8 | via allemaal lijnen | across all those cables | | |
| | die daar gespannen zijn | that are spanned there | | |

of parasocial responses can have. The listeners in the corpus respond to “mushrooms”, while the parasocial responses are in reaction to the refining statement “500 grams”. We have seen in Section 3.1.3 that the parasocial responses are on average 220 ms slower. Since the pause between the two statements is very short (a little over 100 ms), this delay would cause the person to place the parasocial response during the “500 grams” statement. Instead they wait until the refining statement is finished. However, the faster responses from the listeners do not interfere with this statement and are made before the refining statement is started. Response opportunity 7 is a *controversial* moment since it is also yucked by two individuals. This is probably due to the timing, which is synchronous to the start of the word “and”.

Besides response opportunities 4 and 7 there are other *controversial* response opportunities in the corpus. The most controversial moment has 6 positive responses (3 MultiLis and 3 parasocial) and 3 negative yuck responses. The transcript of this moment is presented in Table 5.2. In this segment the speaker corrects himself. An acknowledgment from the listener through a listener response is valid according to six perspectives. The recorded listeners all responded to this moment, however two of them did not respond with a head nod, but with a polite smile (the speaker also smiles at this moment). However, the generated virtual agent in study 3 only performs a head nod. So it is likely that the response opportunity is not yucked because of the timing, but because of the type of listener response displayed.

Another reason for controversy in the corpus is that two response opportunities in quick succession (within 2 seconds) are individually regarded as good response opportunities (at least 4 positive responses to each opportunity in the first two studies), but when generating a listener response at both moments in the third study, the second listener response gets yucked by some individuals.

The last category of responses are the *neutral* responses. These are responses which are generated as *between-head-nods* in Study 3 at moments they received no positive responses in the first two studies. However, in the third study they were not seen as inappropriate responses and thus not yucked. In Table 5.3 one of these moments is transcribed. The head nod is placed mid-sentence, not during a pause. The complete statement is not yet finished. However, it is placed directly after a vital piece of information within this statement (“an opposing building”), which is emphasized by the speaker and memorized after a short hesitation. A confirmation of this piece of information is appropriate according to Study 3 even though no other perspectives previously provided a response there. There are 7 *neutral* moments in our data (see Figure 3.5). In 5 of these moments the listener response is placed mid-sentence after a vital piece of information as in the previous example. In the other two cases the listener response is placed between sentences.

## 5.2 Speech Activity and Pause

As was shown in the previous section, response opportunities are usually placed in response to content that needs to be confirmed and/or assessed. This is in line with the research of Dittman and Llewellyn [40]. They have identified that listener responses are placed near or after the end of a grammatical clause. It is this clause the

listener responses refer to as being heard, understood, agreed upon and/or otherwise assessed. However, an exact time frame within which these listener responses need to be placed is not stated. The question remains whether the listener waits until the speaker has finished the grammatical clause, and if so for how long. Sacks *et al*. [129] suggested that this is not always the case. They said that in interaction interlocutors usually predict the ending of a sentence or turn to plan their response and based on that prediction often respond before the speaker has completed their contribution.

Seeing the number of listener response prediction models that classify interpausal units to be either followed by a listener response or not [114, 28, 136, 88, 133], it is often assumed that listener responses are only placed during pauses. However, there are several findings that contradict this assumption.

In the Dutch IFADV corpus Truong *et al*. [141] observe that only 37% and 16% of the visual and vocal listener responses respectively occur during a pause. In another corpus of a spontaneous dialogue between two Swedish female speakers 40 out of 75 listener responses occurred completely or partially during a pause [101]. These results suggest that at most half the listener responses are placed during a pause.

The fact that listener responses are placed near the completion of a grammatical clause was supported by Truong *et al*. [141], as they note that the probability of listener response increases as speech progresses. They observe a peak in speech probability around 1.25 seconds before the onset of a listener responses, so this seems to be the maximum delay between the grammatical clause and the accompanying listener response.

In the following section analysis will be presented of the relation between speech activity/pause and listener responses in our Dutch MultiLis corpus.

### 5.2.1 Procedure

The analysis of the relation between the occurrence of a listener response and pause used the 1733 response opportunities identified by combining the perspectives from the three recorded listeners. The surrounding seconds of the onsets of these response opportunities were of interest. The number of times the speaker was silent was counted for each frame (sampling rate 100 Hz) from 5 seconds before until 5 seconds after the onset of the response opportunity. This analysis was performed four times; once for all response opportunities and once for response opportunities with only one response (RO1), with two responses (RO2) and three responses (RO3).

Speech labels were automatically extracted using the Dutch ASR software SHoUT [81]. When the gap between two segments was less than 100ms, the two speech segments were combined. Based on these labels the speakers are silent 32% of the time. This number is used as a baseline. If there is a relation between the presence or absence of speech from the speaker and the onset of a response opportunity either a higher or a lower percentage of pause than the baseline of 32% in the region before and after the onset of the response opportunities can be expected.

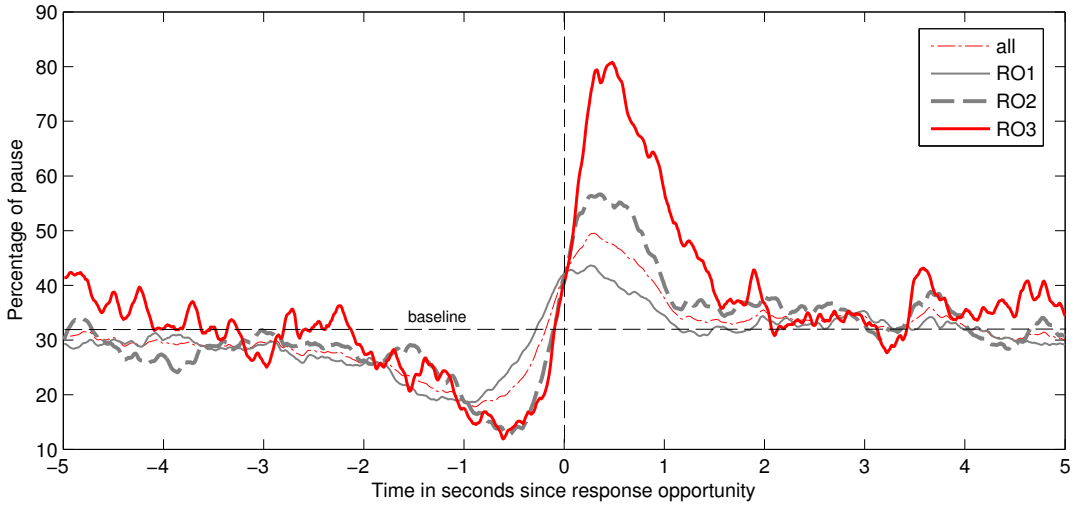


**Figure 5.3:** Correlogram illustrating the percentage of speaker pause around each of the response opportunities. The response opportunities are divided into three groups, one with responses from one listener (RO1), one with responses from two listeners (RO2) and one with responses from three listeners (RO3). The baseline is at 32%, the amount of time the speaker is silent in the MultiLis corpus.

### 5.2.2 Results

The resulting correlogram of the analysis of the relation between pause and listener responses is presented in Figure 5.3. On the horizontal axis time relative to the onset of the response opportuntity is respresented in seconds, starting from 5 seconds before up until 5 seconds after. On the vertical axis the percentage of times the speaker was silent at that time relative to the onset of the response opportunity. The peaks and valleys in the graph show the moments in time relative to the onset of the response opportunity where pause is more or less probable.

At time 0, the onset of the response opportunity, the difference in pause rates between the three different groups of response opportunities are non-existent, with pause rates of 42% for all three groups, which is significantly above our baseline, $\chi^2(1, N = 3470) = 39.0, p < 0.01$. However, there are clear differences between the three groups in the seconds before and after the onset of the response opportunities. For RO2 and RO3 the second before the onset usually contains speech, with pause rates at 0.7 seconds before the onset of 12%. The pause rate for RO1 at that time is 25%. Starting from 0.3 seconds before the onset these rates start to increase massively for RO2 and RO3. Eventually RO2 peaks 0.3 seconds after the onset at a pause rate of 55%, while RO3 peaks a little later at 0.5 seconds after the onset at 81%. RO1 reaches a peak of 43% at 0.3 seconds after the onset. RO1 and RO2 stabilize around the baseline at 1 second after the onset, while RO3 stabilizes 1.5 seconds after the onset.

## 5.3 Energy

The speech signal holds more information than the presence or absence of speech. One of the features of the speech signal is the energy. This is a feature which correlates with the loudness of the speech.

The relation between this energy feature and the timing of listener responses has been established for several languages. A positive correlation was found in Japanese for a high peak and a late decrease in energy in the speech preceeding a listener response and a negative correlation was found for a low peak in energy and a flat energy contour [89]. For English a correlation was observed with high energy [63], but also low energy [99]. For Spanish the low energy cue was also observed [99, 152], while for Chinese a falling energy contour has been related to listener responses [152].

In this section we will analyze what the relation between this feature and listener responses is in our Dutch MultiLis corpus. Since we have already observed that listener responses are placed near the end of utterances, we expect the energy to be high a few hundred milliseconds before the listener response and decrease as we get closer to the onset of the listener response.

### 5.3.1 Procedure

The analysis of the energy values of the speech preceeding listener responses in our Dutch MultiLis corpus will be performed by plotting bitmap clusters [45] following the visualization improvements suggested by Ward and McCartney [151, 152].

The raw energy values were extracted using the openEAR toolkit [46] at a sampling rate of 100 Hz. To normalize for each speaker, all energy values were converted to percentile based values with the heighest valid energy value per speaker being 100% and the lowest value 1%. This conversion is done such that each percentile energy value occurs approximately 1% of the time per speaker.

Using these energy values four bitmap clusters were created, one for all 1733 response opportunities, and one for RO1, RO2 and RO3. The region of interest of these bitmap clusters is the two seconds preceeding the onset of the response opportunity.

In this region of interest the number of times an energy value of a certain percentile at a certain time frame occurs is counted for each response opportunity. So, each point in the bitmap represents the number of times that percentile energy value occurred at that time. The coloring of the bitmap cluster is normalized so that the highest count was black and the lowest white. As a final step the bitmap cluster was smoothed by averaging the values with a 5 by 5 window.

### 5.3.2 Results

In Figure 5.4 the bitmap cluster of the energy values in the two seconds before the onset of all response opportunities is presented. On the horizontal axis time relative to the onset of the response opportunity is represented in seconds, with 2 seconds before the onset on the left end and the time of the onset of the response opportunity on the right end. On the vertical axis the percentile energy is represented with 100% at the top (high energy) and 1% at the bottom (low energy).

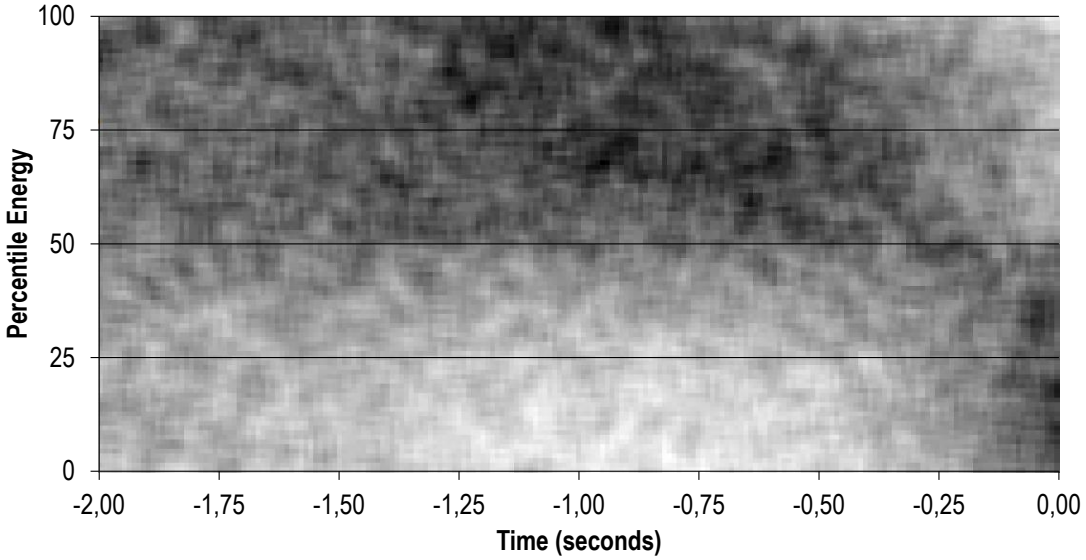


**Figure 5.4:** Bitmap cluster of the percentile energy values in the two seconds preceeding the onset of all 1733 response opportunities.

Looking at the darker areas the bitmap cluster highlights at what time, what percentile energy values occur most frequently. The lighter areas highlight percentile energy values that are usually not seen at that time before a response opportunity.

The bitmap cluster shows that up until 0.5 seconds before the response opportunity the energy is high, above 50%. From there on, the intensity seems to decline, with the switch around 0.25 seconds before the response opportunity, where the majority of the energy becomes low, below 50%. This correlates with the fast increase in pause percentage we have seen in Figure 5.3. From the bitmap cluster it is hard to see whether the intensity level is rising or falling right before the speaker falls silent. This bitmap cluster is similar to the bitmap clusters for Iraqi Arabic, Spanish, American English and to a lesser extent, Chinese [151].

In Figures 5.5 to 5.7 the bitmap clusters for the three types of response opportunity are presented. These figures show that the time of the drop in energy is closer to the onset of the response opportunity as the number of listeners that responded increases. For RO1 this drop starts from 0.5 seconds before the onset. For RO2 this drop occurs 0.25 seconds before. Finally, for RO3 only 0.1 seconds before, if this drop occurs at all. Due to the lower sample size the bitmap cluster for RO3 is less clear than the other two.

The interaction between pause and listener responses we observed in Figures 5.3 and 5.4 is that listener responses usually (closely) follow speech and are situated right before or right at the start of a pause. This interaction is stronger in the cases where more than one listener provides a response (RO2 and RO3). It is interesting to note that the pauses after the response seem to be more frequent and longer in the case of RO3. As noted in Section 5.1 these response opportunities often follow summarizing or refining statements by the speaker. These statements are targeted to improve the understanding of the content by the listener and therefore a listener response - a signal to communicate understanding - can be expected by the speaker. To allow this

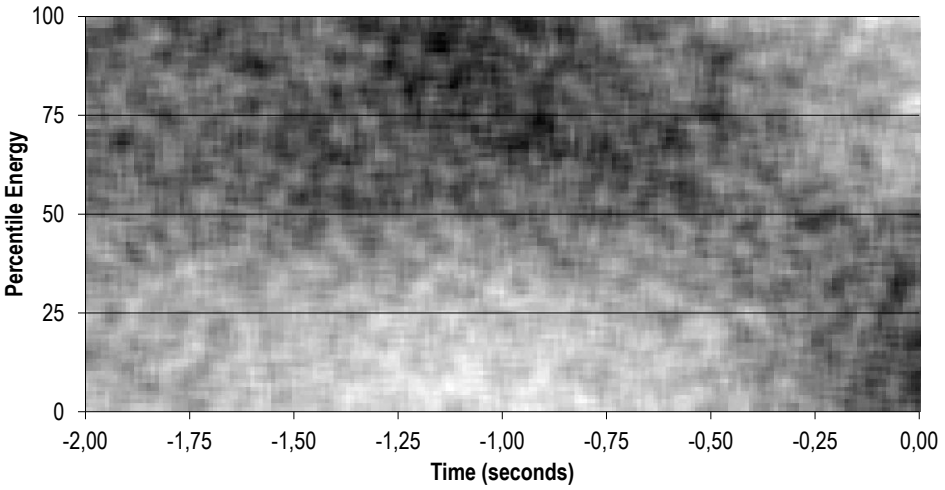


**Figure 5.5:** Bitmap cluster of the percentile energy values in the two seconds preceeding the onset of all 1140 response opportunities with one response (RO1).

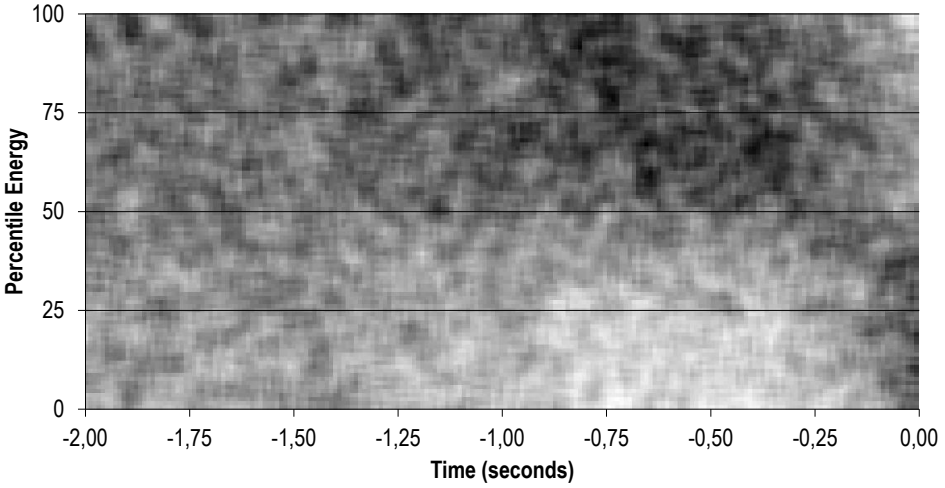


**Figure 5.6:** Bitmap cluster of the percentile energy values in the two seconds preceeding the onset of all 465 response opportunities with two response (RO2).

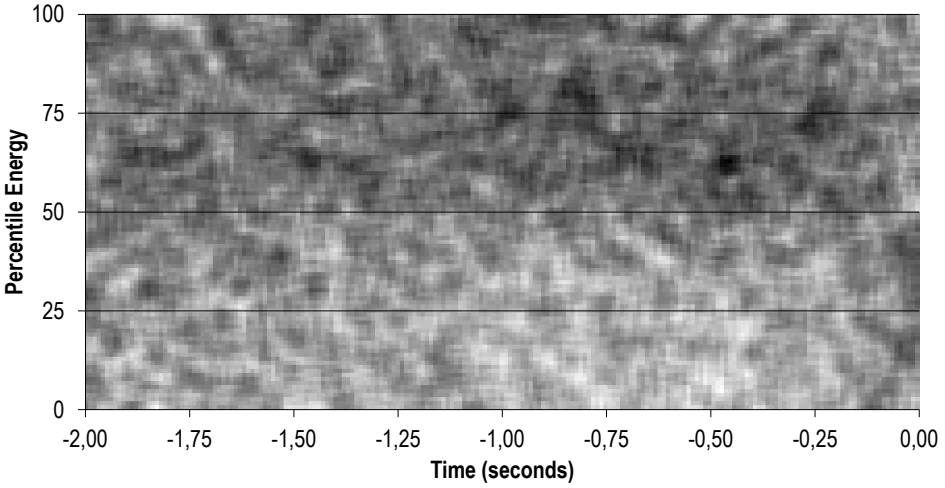


**Figure 5.7:** Bitmap cluster of the percentile energy values in the two seconds preceeding the onset of all 128 response opportunities with three response (RO3).

to happen, the speaker extends his or her pause slightly. This type of coordination between speaker and listener is studied and described in more detail in [57, 34, 7, 9].

## 5.4 Pitch

A prosodic feature that is often related to the timing of listener response is the pitch or fundamental frequency of the speech. Several analyses of the relation between the pitch cue and listener responses have been performed on manually annotated data. Coders have manually annotated the contour of the pitch of the speech. These contours are descriptions of the change and value of the pitch in a speech segment and are expressed in terms of rising, falling, flat/sustained and high, low or medium pitch.

Based on such annotations, Dittmann and Llewellyn [40] reported that in an English corpus a listener response more often follows a speech segment with a falling or rising pitch contour as opposed to a speech segment with a sustained pitch contour. A speech segment with a falling or rising pitch contour is followed by a response 37.1% or 26.4% respectively, while a speech segment with a sustained slope is followed by a response only 6.1% of the time. Gravano and Hirschberg [63] reported that the final 200 to 300 ms of the interpausal unit preceeding a listener response has a rising contour and is generally higher in pitch. For French, Bertrand *et al*. [13] observed that the rising or sustained contours are relevant at points where listener responses occur.

Analyses of the pitch cue have also been performed based on automatically extracted pitch contours on several languages. The pitch contour of the speech preceeding a listener response has been described as rising for Japanese [89, 114] and Swedish [133]; rise-falling for Japanese [89, 114]; falling for Arabic [149, 150] and Swedish [133]; high for Swedish [133] and low for English [153, 99, 151], Chinese [152], Japanese [151, 152], Swedish [133], Dutch [141] and Spanish [99].

### 5.4.1 Procedure

The analysis of the pitch contours preceeding listener responses in the MultiLis corpus was performed by plotting bitmap clusters [45] following the visualization improvements suggested by Ward and McCartney [151, 152].

The raw pitch features were extracted using the algorithm from Drugman and Alwan [42] at a sampling rate of 100 Hz. Our interest is only in the pitch values of speech, so only the pitch values during the speech segments as detected by the ASR software SHoUT [81] are considered valid.

To be able to calculate the pitch of speech a sound is required whose frequency is clear and stable enough to distinguish from noise. This is not true for all segments of speech. Most consonants do not have a harmonic frequency spectrum and this makes automatic detection of pitch for these consonants near impossible. This results in segments of speech where no pitch value can be established. To deal with this fact, gaps in detected pitch smaller than 80 ms (8 frames) are linearly interpolated, following [153]. Finally, all pitch values are converted to percentile based values with the heighest valid pitch value per speaker being 100% and the lowest value 1%. This

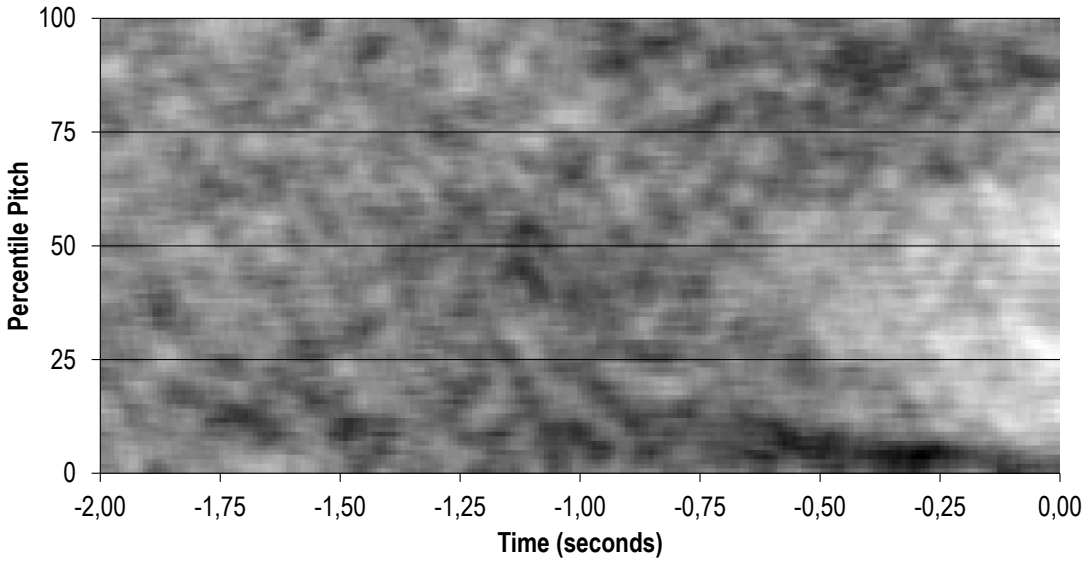


**Figure 5.8:** Bitmap cluster of the percentile pitch values in the two seconds preceeding the onset of all 1733 response opportunities.

conversion is done such that each percentile pitch value occurs approximately 1% of the time per speaker.

Using these pitch values, four bitmap clusters are created, one for all 1733 response opportunities, one for the 1140 response opportunities with only one response (RO1), one for the 465 response opportunities with two responses (RO2) and one for the 128 response opportunities three responses (RO3). The region of interest of these bitmap clusters is the two seconds preceeding the onset of the response opportunity.

In this region of interest for each response opportunity the number of times a pitch value of a certain percentile occurs at a certain time frame was counted. So, each point in the bitmap represents the number of times that percentile pitch value occured at that time. The bitmap cluster was normalized so that the highest count is black and the lowest white. As a final step the bitmap cluster was smoothed by averaging the values with a 5 by 5 window.

### 5.4.2 Results

In Figure 5.8 the bitmap cluster of all 1733 response opportunities found the MultiLis corpus is shown. On the horizontal axis time relative to the onset of the response opportunity is represented in seconds, with 2 seconds before the onset on the left end and the time of the onset of the response opportunity on the right end. On the vertical axis the percentile pitch is represented with 100% at the top (high pitch) and 1% at the bottom (low pitch).

Looking at the darker areas the bitmap cluster highlights at what time, which percentile pitch values occur most frequently. The lighter areas highlight percentile pitch values that are not usually seen at that time before a response opportunity.

In Figure 5.8 two trends can be seen. The pitch either becomes very low (below the 10th percentile mark) in the few hundred milliseconds preceeding the response opportunity or high (above the 60th percentile mark). So, the pitch values between

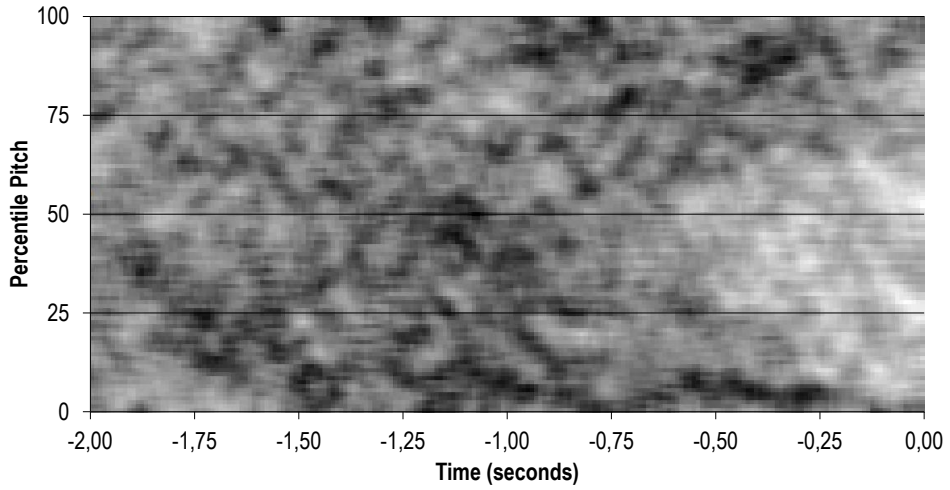


**Figure 5.9:** Bitmap cluster of the percentile pitch values in the two seconds preceeding the onset of all 1140 response opportunities with one response (RO1).

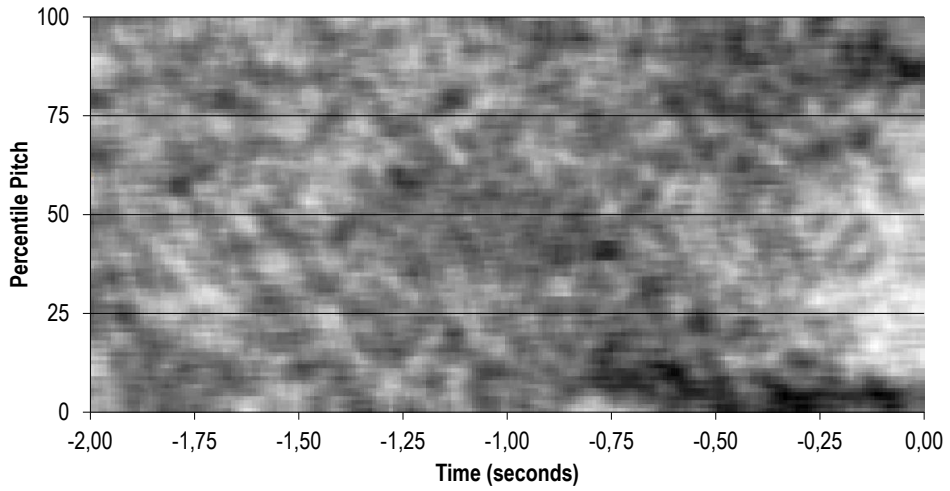


**Figure 5.10:** Bitmap cluster of the percentile pitch values in the two seconds preceeding the onset of all 465 response opportunities with two responses (RO2).

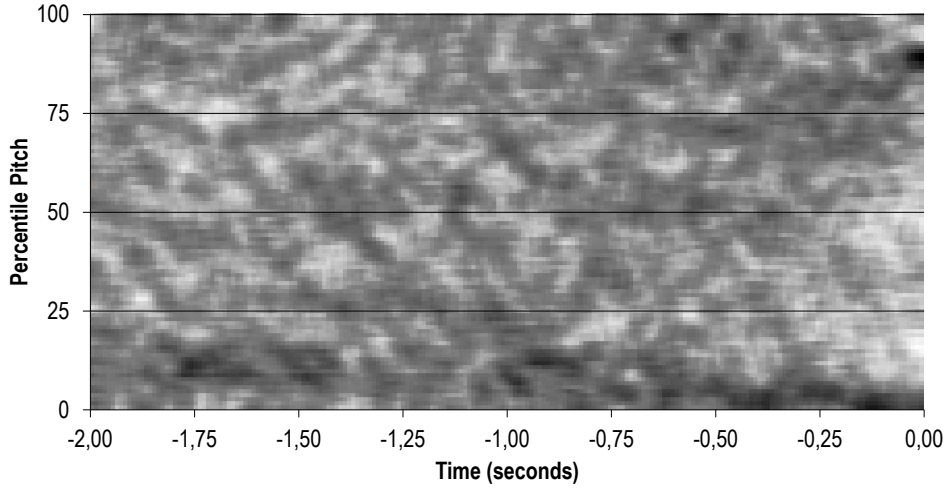


**Figure 5.11:** Bitmap cluster of the percentile pitch values in the two seconds preceeding the onset of all 128 response opportunities with three responses (RO3).

the 10th and 60th percentile mark occur less frequently than one would expect. This divergence seems to emerge starting around 750 ms before the onset of the response opportunities. Any pitch values before this time do not seem to have a correlation with the occurrence of a response opportunity.

A bitmap cluster was created for each of the three types of response opportunity to see whether there are differences in the pitch values preceeding a response opportunity with a different number of listeners responding. Figure 5.9, Figure 5.10 and Figure 5.11 show these bitmap clusters of two seconds preceeding the response opportunities with one (RO1), two (RO2) and three (RO3) responses, respectively.

The bitmap cluster for RO1 (Figure 5.9) looks very similar to the original bitmap cluster (Figure 5.8). This is not surprising since this bitmap cluster is based on 66% of the same response opportunities. Again the same patterns as described above can be seen. In the bitmap cluster for RO2 (Figure 5.10) the very low pitch cue (below 10%) seems to be present more often for response opportunities with two responses than for response opportunities with one response. A very high pitch (above 75%) in the final 250 ms can also be observed. The bitmap cluster for RO3 (Figure 5.10) is less outspoken because of the lower number of samples (128). However, the very low pitch value cue can still be observed and a very high pitch in the final 100 ms before a response opportunity with three responses.

These results confirm the earlier findings that in most languages, including Dutch [141], a listener response is preceeded by a region of low pitch [153, 99, 151, 152, 133]. Furthermore, a region of high pitch is also observed. Although extracting a contour from these bitmap clusters is not arbitrary, the bitmap clusters suggest that the pitch begins rising or falling to these high and low values, respectively, around 750 ms before the listener response.

## 5.5 Eye Gaze

Bavelas *et al*. [8] studied the relation between gaze and listener responses through microanalysis of their corpus. In this research they identified the collaboration between the speaker and the listener, where the speaker typically seeks a response from the listener by looking at him or her, which begins a brief period of mutual gaze. When the listener responds within this so-called *gaze window*, the speaker quickly looks away, terminating the window and continuing to hold the turn. These findings have confirmed earlier findings by Kendon [86] and Argyle *et al*. [6] stating that a speaker looks at his/her interlocutor at points in the discourse where the speaker is looking for a response from his/her interlocutor.

This relation has been found in several corpora. Bavelas *et al*. [8] observe that 83% of the listener responses are given when the speaker looks at the listener. Truong *et al*. [141] observe a percentage of 85% on the Dutch IFADV corpus.

For the recreation of this behavior in a virtual human more information is needed than these studies provided. The previous studies have given general descriptions of the behavior, but do not give us the specifics on the timing of the response once a gaze window has been established. How long into the gaze window does the listener typically respond? And how long does the gaze window stay open after the response

has been given?

### 5.5.1 Procedure

The relation between gaze and listener responses in the MultiLis corpus was analyzed. Speaker gaze was manually annotated. Unfortunately, annotations for gaze of the listener were not available, so the relation between the gaze window discussed by Bavelas *et al*. [8] and listener responses could not be analyzed; only the relation with speaker gaze. Since the listener usually looks often and for long periods in time at the speaker [5, 86, 141], results can be expected to closely resemble results on mutual gaze data.

The 1733 response opportunities identified by combining the perspectives from the three recorded listeners were used for the analysis. In this analysis we are interested in the seconds surrounding the onset of the response opportunities. To this end the number of times the speaker was looking at the listener were counted for each frame at a 100 Hz sampling rate from 5 seconds before until 5 seconds after the onset of the response opportunity. This analysis was performed for all response opportunities, for response opportunities with only one response (RO1), with two responses (RO2) and three responses (RO3).

The baseline was established by the total percentage of the speaker looking at the listener in our corpus. If there is a relation between the presence or absence of gaze from the speaker and the onset of a response opportunity, an either higher or lower percentage of gaze than the baseline of 67% in the region before and after the onset of the response opportunities can be expected.

### 5.5.2 Results

The resulting correlogram of the procedure discussed above is presented in Figure 5.12.

By looking at time 0 one can see that for response opportunities with one, two and three responses the gaze rates are 81%, 90% and 95%, respectively. These differences between RO1, RO2 and RO3 are significant, $\chi^2(1, N = 1735) = 30.6, p < 0.01$. By looking at time -2 to 0 in the graph one can observe that the gaze rate starts to increase about 2 seconds before the actual response opportunity. Starting from 1 second before the response opportunity RO3 has a higher gaze rate than RO2 and RO1. The gaze rate sharply declines during the 1 second after the response has been given for all three categories, where RO3 continues to drop a little further.

These observations illustrate the collaboration between speaker and listener identified by Bavelas *et al*. [8]. Indeed, the speaker seeks a response from the listener by looking at him or her and after this response has been given the speaker looks away. These results suggest that this interaction takes about three seconds, where the speaker usually looks at the listener for up to two seconds before the listener responds and continues to look at the listener for up to one second afterwards.

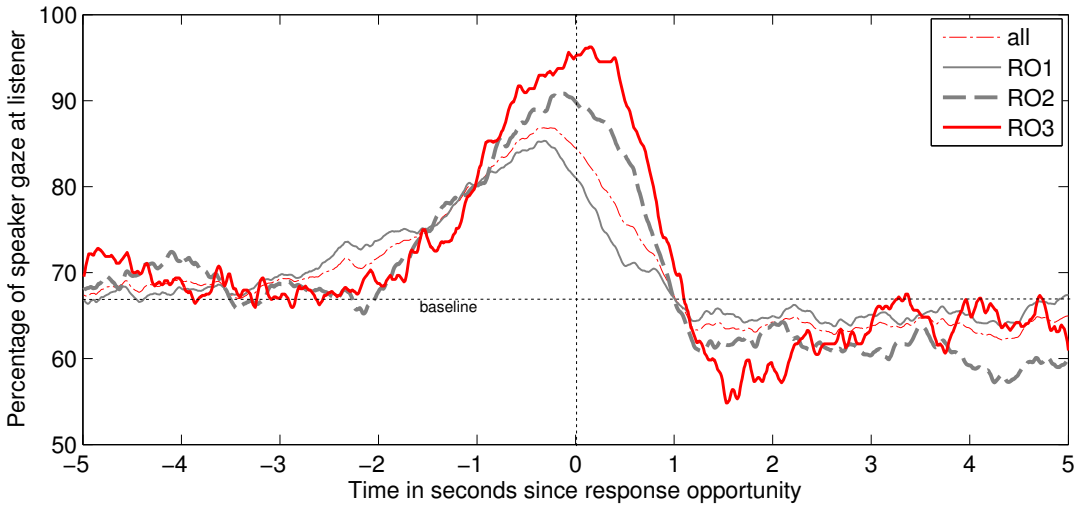


**Figure 5.12:** Correlogram illustrating the percentage of gaze from the speaker at the listener around each of the response opportunities. The response opportunities are divided into three groups, one with responses from one listener (RO1), one with responses from two listeners (RO2) and one with responses from three listeners (RO3). The baseline is at 67%, the amount of time the speaker looks at the listener in the MultiLis corpus.

## 5.6 Head Gesture

The final part of this chapter will take a look at the form of the listener response. Whether the different listeners give a listener response of the same form when they respond to the same response opportunity or not will be analyzed.

### 5.6.1 Agreement in Head Gesture

As explained in Section 2.2.2, the main head gesture types available in the MultiLis corpus are Nod (characterized by a downward stroke), Backnod (characterized by an upward stroke) and Double nod (two nods in quick succession with the same amplitude). There is a Lingering variant of each of these head gesture types. These head gestures continue for a period of time in decreasing amplitude. If the amplitude is increased a new head gesture is annotated. The corpus also included the labels Downstroke and Upstroke, which are single up or down movements. For the next analyses these are considered as Nod and Backnod, respectively. All the other labels are combined in the label Other.

The first analysis looks to see if the three listeners used the same head gesture when they reacted at the same time. For each response opportunity with at least two responses we noted the head gestures type the listener used in their response. This was Nod, Backnod, Double Nod or Other. Then Krippendorff's alpha coefficient [94] was calculated to measure the agreement between the two or three listeners, resulting in an alpha of 0.082. So, no significant agreement was found in the head gesture type the listeners used.

Agreement between listeners in their use of the lingering head gestures was also analyzed. Each head gesture was coded as either Lingering or Not Lingering. Again

Krippendorff's alpha coefficient was calculated to measure agreement and no agreement was found (alpha = 0.017).

### 5.6.2 Typical vs Deviated Responses

So, the results of the annotations in the MultiLis corpus show that listeners do not all produce a head nod starting downwards, a head nod starting upwards nor two similar nods in quick succession in response to the same response opportunity. However, this study did not control for the individual differences in listener response production. Looking at the listening behavior in the MultiLis corpus shows that each individual has a typical head nod they usually produce. One individual may usually start their head movement downward, whereas another usually starts upwards. Given this knowledge, it is not surprising that the previous analysis showed no agreement.

However, there are times listeners deviate from their typical listener response. This deviation could either be a bigger amplitude, a higher frequency, louder voice and/or different form. These responses seem to bear more meaning than these typical responses. The listeners seem more sure that the response needs to be given and show this by giving a clearer signal to the speaker.

A distinction can thus be made between *typical* and *deviated* responses. In this distinction the typical responses are the responses that, based on the form of the response, a listener opts for when no meaning is attached to the response other than being a signal of attendance and superficial understanding; the response that the listener will give the majority of the response opportunities (at least in our data). The deviated responses are the responses that seem to bear more meaning as they deviated from the typical response in terms of increased amplitude, velocity and/or frequency of the head gestures or are different in form altogether.

This distinction is closely related to a distinction in listener responses made by Bavelas *et al*. [8]. They made the distinction between generic and specific listener responses (or continuers and assessments as Goodwin calls them [57]). In this distinction generic listener responses keep the listener clearly in the role of listener/audience, whereas specific responses are co-telling acts in which the listener becomes involved with the narrator telling the story.

Our interest in the division was the form of the listener response, not its relation to the speaker's actions. Since annotations were performed without the audiochannel of the speaker, an accurate assessment of whether a response was co-telling or not could not be achieved solely on the form basis.

Despite the different viewpoints on which the distinctions between the types of listener responses are based, they are likely to be similar. The typical responses can be assumed to be mostly generic listener responses and it is likely that all specific listeners response are classified as deviated responses. When a listener wants to show his/her involvement with the story, he/she needs to deviate from his/her usual behavior to signal this to the speaker. The deviated responses class also contains "firm" generic responses as well. There are instances where a clear signal is given that what has just been said is understood and that the story or recipe can progress. This is usually achieved by giving a head nod with a larger amplitude, velocity and/or frequency than typical for the listener. Such responses would still be considered a generic

| | n | Typical Response | Deviated Response | % |
|---|---|---|---|---|
| **RO1** | 1142 | 858 | 284 | 24,9% |
| **RO2** | 439 | 598 | 280 | 31,9% |
| **RO3** | 119 | 229 | 128 | 35,9% |
| **Total** | 1700 | 1685 | 692 | 29,1% |

**Table 5.4:** Correlation between the number of listeners responding to a response opportunity (one listener = RO1, two listeners = RO2, three listeners = RO3) and whether they produced a typical or deviated response.

response by the division of Bavelas *et al*.

The following analysis will test whether a correlation exists between these typical and deviated responses and the number of listeners that responded to a response opportunity. The results of the previous analyses have shown that response opportunities with more than one response are often more clearly cued by the speakers. At those moments speakers often made a more significant contribution to the discourse and looked at the listener more often. Since there are more cues given by the speaker at those moments, the listeners are more aware that the speaker wants a response at this moment and may opt for a deviated response more often than a typical response to acknowledge this fact.

### Procedure

To collect annotations one annotator classified 2377 listener responses from the MultiLis corpus as either a *typical* response or a *deviated* response. What the typical response was for each individual listener was established by watching each of the three videos of the particular listener before actually classifying any responses. The prevalent response was regarded to be the typical response for that particular listener. Sometimes the listener has a small set of prevalent responses the listener alternates. If so, all of these responses were annotated as a typical response. For this annotation only responses with a head and/or vocal component were considered. Thus, responses only consisting of smiles, raised or frowned eyebrows and/or other gestures were not considered in this study. Annotation was performed with no sound, solely on the video image. A second annotator classified 8 out of 32 listeners and inter-annotator agreement between the two annotators was 81% with a Krippendorff's alpha of 0.56.

### Results

In Table 5.4 the results of the annotations are presented. The 2377 response are divided over 1700 response opportunities[1]. From the 2377 responses 1685 responses were classified as typical responses and 692 as deviated responses for a deviation rate of 29.1%. Splitting up the responses of the three different type of response opportunities shows that deviation rate increases from 24,9% for RO1 to 31,9% for RO2

[1]Note that the numbers of RO1, RO2 and RO3 deviate slightly from the previous analyses due to the fact that only listener responses with a head and/or vocal component were considered

and 35,9% for RO3. A $\chi^2$ test shows that this increase is significant ($\chi^2(2, 2377) = 21.1, p < 0.001$). Thus, at the moments where more than one listener responds (RO2 and RO3) it is significantly more likely that a listener will choose a deviated response.

## 5.7 Conclusion

The conversational analyses presented in this chapter gave insights into the relation between actions of the speaker and the occurrence of a response opportunity. With the study on the IFADV corpus [141], this is the second corpus on which listener response inviting cues have been investigated for Dutch. The findings mostly corroborate the findings on the IFADV corpus. For eye gaze and speech activity the analyses mostly findings from earlier literature on other languages as well. The pitch correlogram mostly resembles the correlograms for Spanish and English [151], but the high pitch cue is much more present is this corpus.

The novel aspect of these analysis is that the MultiLis corpus gives insight into the graded optionality of the response opportunities. The correlograms for eye gaze and speech activity showed that these cues are more often present for response opportunities where all three listener responded than for response opportunities where only two or one listener responded. This suggests that the speaker used more cues to signal the response opportunity at these moments and the listener responded to that. Furthermore, the setup of the MultiLis corpus allowed us to compare the chosen form of the listener response from different listeners. The results showed that listeners are more likely to deviate from their typical response when the response opportunity is less optional (more listener responded). This suggests that the listeners recognize that the speaker specifically wants a listener response to be given at these moments and makes an extra effort to give one.

With these analyses we identified several behavioral patterns of the speakers that are useful for the automatic prediction of the timing of listener responses, which will be the topic for Part III of the thesis. We have seen that features describing the speech activity, energy, pitch and eye gaze of the speaker all contain cues for listener responses. Thus, these will be used as input features for the models that are learned on the MultiLis corpus.

# Part III

# Predicting the Timing of Listener Responses

# 6

# Listener Response Prediction Models

This Chapter will give a survey of the listener response prediction models that have been developed over the years. The survey will focus on listener response prediction models that have been developed based on observations on human-human interactions or are directly learned from these interactions. Another requirement for inclusion was that the performance of the model has been evaluated.

Since 1989, when Watanabe & Yuuki [154] proposed their "voice reaction system with a visualized response equivalent with nodding", a wide variety of listener response prediction models have been proposed. These models have mainly focused on the prediction of generic listener responses. Models for specific listener responses and closely related phenomena such as grounding and detection of turn relevant places are out of the scope of this survey.

Even though the proposed prediction models for generic listener responses differ in many aspects, the general approach to the problem has not changed much since the first model by Watanabe & Yuuki. Based on a corpus of human-human interactions a model is learned that matches the behavior of the speaker to the presence or absence of a generic listener response from the listener. These models for generic listener response react to easily extractable features from audio and (sometimes) video signal, describing the behavior of the speaker. Based on these features the model infers a decision value that is associated with the occurrence of a generic listener response. If this decision value is high enough a listener response is predicted. The predictions made by the model are then evaluated.

What follows is a survey of the listener response prediction models that have been proposed in the past. An overview of the models included in the survey is presented in Table 6.1. The survey focusses on the differences between approaches. These differences will be discussed for the following aspects.

- **Corpus** - During the development and evaluation of the listener response prediction model a corpus containing examples of human-human interaction is used. In these corpora the behavior of the speaker and listener are annotated. The corpora that are used differ in language, available communication channels between interlocutors, topic and nature of the interaction. (Section 6.1)

- **Features** - The behavior of the speaker is described in a feature vector that is constructed by extracting descriptive features from the audio and/or video signal of the recordings. Proposed approaches differ in which features are used and how features from multiple modalities are combined. (Section 6.2)

- **Model** - The model is the technique that is used to infer the relation between the feature vector with observations describing the behavior of the speaker and the occurence of a listener response in the corpus. The proposed approaches differ in segmentation of the data and machine learning model used. (Section 6.3)

- **Evaluation** - In the end the model is evaluated. This is done by comparing the generated behavior based on the predictions of the model to the behavior of the recorded listener and/or by having human judges rate the generated behavior. Besides this distinction approaches also differ in performance measure, margin of errors and ground truth selection. (Section 6.4)

## 6.1 Corpus Data

The corpora used in the development and evaluation of the listener response prediction models differed in several aspects, such as language, communication channels, topic of the interaction and distribution of the speaking role.

The first models were developed on corpora in the Japanese language [154, 115, 114, 153, 51, 136, 88, 113]. Later models were developed using corpora in other languages, such as English [153, 28, 107, 79, 78, 108, 116, 118, 124, 110, 100, 117], Chinese [151], Spanish [151, 100], Arabic [100], Dutch [121] and Swedish [133].

The corpora also differed in the communication channels that were available to the interlocutors. The interactions were either only using the audiochannel, so the interlocutors could not see each other [115, 153, 136, 88, 113, 151] or face-to-face [154, 114, 51, 107, 79, 78, 108, 116, 118, 100, 117]. The HCRC Map Task corpus used in [28, 110] includes interactions in both settings. Poppe *et al*. [121] used avatar-mediated interactions, where the speaker was talking to a generated virtual listener. Skantze [133] used interactions where the speaker had to imagine the presence of a listener.

Finally, the topic and nature of the interaction differed between the corpora. In some corpora the interlocutors were assigned roles, where one interlocutor was assigned the role of speaker and the other the role of listener [154, 115, 28, 107, 79, 78, 108, 116, 118, 124, 110, 100, 121, 133, 117], while in others no such distinction was made in advance [114, 153, 136, 51, 88, 113, 151]. The topics of conversation ranged from free topic dialogue [114, 153, 136, 51, 88, 113, 151] or monologue [121] and emotionally colored dialogues [124] to more structured interaction such as the uttering of a telephone message [154], telephone ordering [115], summarizing a video clip [107, 79, 78, 108, 116, 118, 117], storytelling [100] and giving navigation instructions [28, 110, 133].

| | Corpus | | | Features | | | | | Model | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Authors** | **Language** | **Modality** | **Topic** | **Acoustic** | **Part-of-Speech** | **Lexical** | **Syntactic** | **Visual** | **Segmentation** | **Technique** |
| Watanabe & Yuuki (1989) [154] | Jap | F2F | Telephone message | ✓ | | | | | Continuous | Handcrafted |
| Okato *et al.* (1996) [115] | Jap | Audio | Telephone ordering | ✓ | | | | | Continuous | HMM |
| Noguchi *et al.* (1998) [114] | Jap | F2F | Free topic | ✓ | | | | | Inter-pausal unit | Decision Tree |
| Ward & Tsukahara (2000) [153] | Jap/Eng | Audio | Free topic | ✓ | | | | | Continuous | Handcrafted |
| Cathcart (2003) [28] | Eng | F2F/Audio | Map Task | ✓ | ✓ | | | | Words | Handcrafted |
| Fujie *et al.* (2004) [51] | Jap | F2F | Free topic | ✓ | | | | | Continuous | FST |
| Takeuchi (2004) [136] | Jap | Audio | Free topic | ✓ | ✓ | | | | Pause-frames | Decision Tree |
| Kitaoka *et al.* (2005) [88] | Jap | Audio | Free topic | ✓ | ✓ | | | | Pause-frames | Decision Tree |
| Nishimura *et al.* (2007) [88] | Jap | Audio | Free topic | ✓ | ✓ | | | | Pause-frames | Decision Tree |
| Morency *et al.* (2008) [107] | Eng | F2F | Video summary | ✓ | | ✓ | | ✓ | Continuous | HMM/CRF |
| Huang *et al.* (2010) [79] | Eng | F2F | Video summary | ✓ | | ✓ | | ✓ | Continuous | CRF |
| Huang *et al.* (2010) [78] | Eng | F2F | Video summary | ✓ | | ✓ | | ✓ | Continuous | CRF |
| Morency *et al.* (2010) [107] | Eng | F2F | Video summary | ✓ | | ✓ | | ✓ | Continuous | HMM/CRF |
| Ozkan & Morency (2010) [116] | Eng | F2F | Video summary | ✓ | ✓ | ✓ | | ✓ | Continuous | CRF |
| Ozkan *et al.* (2010) [116] | Eng | F2F | Video summary | ✓ | ✓ | ✓ | ✓ | ✓ | Continuous | CRF/LDCRF/LMDE |
| Poppe *et al.* (2010) [124] | Eng | F2F | Emotionally colored | ✓ | | | | ✓ | Continuous | Handcrafted |
| Ward & McCartney (2000) [153] | Chn/Spa | Audio | Free topic | ✓ | | | | | Continuous | Handcrafted |
| Neiberg & Gustafson (2011) [110] | Eng | F2F/Audio | Map Task | ✓ | | | | | Continuous | GMM |
| Levow & Wang (2012) [100] | Eng/Spa/Ara | F2F | Storytelling | ✓ | | | | | Inter-pausal unit | Decision Tree/SVM |
| Poppe *et al.* (2012) [121] | NL | Avatar | Free topic | ✓ | | | | | Continuous | SVM |
| Skantze (2012) [133] | Swe | Imagined | Map Task | ✓ | | | | | Inter-pausal unit | Decision Tree |
| Ozkan & Morency (2013) [117] | Eng | F2F | Video summary | ✓ | ✓ | ✓ | ✓ | ✓ | Continuous | CRF/LDCRF/LMDE |

**Table 6.1:** Overview of the corpus based listener response prediction models developed so far.

## 6.2 Features

The behavior of the speaker the model needs to react to is described by a feature vector. This consists of several features that describe several aspects of the behavior extracted from the audio and/or video signal observing the speaker. Most models only use features extracted from the audio signal, but more recently the video signal has been used as well.

Features that have been extracted from the audio signal are a binary representation of speech/non-speech [154, 115, 114, 153, 28, 136, 88, 113, 107, 79, 78, 108, 116, 118, 124, 151, 110, 100, 133, 117], pitch [115, 114, 153, 136, 51, 113, 107, 108, 116, 118, 124, 151, 110, 100, 121, 133, 117], energy/intensity/power [115, 114, 136, 51, 88, 113, 107, 108, 116, 118, 110, 100, 121, 133, 117], part-of-speech tags [28, 136, 88, 113, 116, 118, 117], individual words [107, 79, 78, 108, 116, 118, 117], grammatical function of the words (e.g. subject, object etc.) [118, 117], MFCC[110, 121] and voice quality [100]. Especially for pitch and energy, not only the raw values are used, but also descriptions of the contour.

The features that have been used from the video signal so far are eye gaze (whether the speaker looks at the listener or not) [107, 79, 78, 108, 116, 118, 124, 117], head movement [118, 117] and eyebrow movement [118, 117].

These features can be categorized in several modalities, such as acoustic (speech/non-speech, pitch, energy, MFCC, voice quality), lexical (words), part-of-speech (tags), syntactic (function of the words) and visual features. These feature categories each work on a different time-scale. When combining these features categories for a multimodal prediction model one of two approaches can be taken; early fusion or late fusion. In early fusion models the multiple modalities are merged in the feature stage and ultimately one model is learned, while in late fusion models separate models are learned for each modality and the outcome of these models is then combined to give the final prediction. The approach by Ozkan *et al*. [118, 117] is currently the only late fusion model developed.

## 6.3 Models

Two types of model; *continuous* models and *segmented* models.

The distinction between these models is in the rate at which the model produces predictions. Continuous models make a prediction at each frame. This frame is a fixed unit of time at which the incoming signals are sampled and processed. The most common frame rates for continuous models are 30 and 100 Hz.

Segmented models only make a prediction after a certain event has taken place in the interaction. These events are the completion of an inter-pausal unit (in other words, the detection of a pause), the completion of a word or each 100 ms of pause.

In the following sections more details about these models and the machine learning techniques that are used for them will be explained.

### 6.3.1 Continuous Models

The majority of models process the incoming data from the speaker continuously [154, 115, 153, 107, 108, 116, 118, 124, 117]. This means that a prediction for a listener response can be made at each frame of the interaction. No specific events are required before a prediction is made by the prediction model.

For this continuous classification task researchers have used handcrafted rules based on corpus statistics [154, 153, 124, 151], Finite State Transductor [51], Hidden Markov Model [115, 107, 108], Support Vector Machines [121], Conditional Random Fields [107, 79, 78, 108, 116, 118, 117], Latent Dynamic Conditional Random Fields [118, 117] and Latent Dynamic Mixture of Experts model [118, 117].

### 6.3.2 Segmented Models

As mentioned earlier, the unit of segmentation for segmented models differs between approaches. Several researchers [114, 110, 100, 133] have used inter-pausal units as segments. These models are only capable of predicting a listener response after/during a pause. These proposed models predict for each such segment whether it is followed by a listener response or not. The segmentation used by Cathcart [28] is at the word level. Both Takeuchi *et al*. [136] and Kitaoka *et al*. [88] have proposed a model that classifies frames with no speech from the speaker. These pauses were split into segments of 100ms. For each of these segments the pause was classified as either 'making a listener response', 'taking the turn', 'waiting for the speaker to continue' or 'waiting to make a listener response or take the turn'.

The machine learning techniques researchers have used for the classification of these segments are several types of decision trees [114, 136, 88, 113, 100, 133], handcrafted rules based on corpus statistics [28], the Gaussian Mixture Model [110] and Support Vector Machines [100, 133].

The models that segment the incoming data into bigger chunks are at a disadvantage when it comes to placement of listener responses. They usually limited placement of listener responses to places where the speaker is silent. However, in Section 5.2 it was shown that only about 50% of listener response are actually placed at such times. Listener responses are often placed a few hundred milliseconds before the end of an utterance.

## 6.4 Evaluation

Finally, the models that are developed need to be evaluated. Also in this area there are many differences between approaches as is shown by the overview in Table 6.2. The main distinction between evaluations is between *objective* and *subjective* evaluations. For objective evaluations the performance is measured by comparing the predictions of the model to the recorded ground truth labels. For subjective evaluations the performance is measured by human judgment of the generated behavior. In the following section more details will be presented about similarities and differences between approaches in this regard.

| Authors | Subjective | Objective | Objective Metric | Ground Truth | Segmentation | Margin of Error |
|---|---|---|---|---|---|---|
| Watanabe & Yuuki (1989) [154] | | ✓ | Cross-Correlation Coef. | Multiple (Nodding) | Continuous | - |
| Okato *et al.* (1996) [115] | | ✓ | Precision / Recall | Single | Continuous | -100/500ms |
| Noguchi *et al.* (1998) [114] | | ✓ | Precision / Recall | Multiple (Keyboard) | Inter-pausal Unit | - |
| Ward & Tsukahara (2000) [153] | | ✓ | Precision / Recall | Single | Continuous | -500/500ms |
| Cathcart (2003) [28] | | ✓ | $F_1$ | Single | Words | - |
| Fujie *et al.* (2004) [51] | ✓ | | - | - | - | - |
| Takeuchi (2004) [136] | ✓ | ✓ | Precision / Recall | Single | 100ms Pause Frames | - |
| Kitaoka *et al.* (2005) [88] | ✓ | ✓ | $F_1$ | Multiple (Keyboard) | 100ms Pause Frames | - |
| Nishimura *et al.* (2007) [113] | ✓ | | - | - | - | - |
| Morency *et al.* (2008) [107] | | ✓ | $F_1$ | Single | Continous | 0/1000ms |
| Huang *et al.* (2010) [79] | ✓ | | - | - | - | - |
| Huang *et al.* (2010) [78] | ✓ | | - | - | - | - |
| Morency *et al.* (2010) [108] | | ✓ | $F_1$ | Single | Continous | 0/1000ms |
| Ozkan & Morency (2010) [116] | | ✓ | $F_1$ | Single | Continous | 0/1000ms |
| Ozkan *et al.* (2010) [118] | | ✓ | $F_1$ | Single | Continous | 0/1000ms |
| Poppe *et al.* (2010) [124] | | ✓ | $F_1$ | Single | Continous | -200/200ms |
| Ward & McCartney (2010) [151] | | ✓ | Precision / Recall | Single | Continuous | -500/500ms |
| Neiberg & Gustafson (2011) [110] | | ✓ | Recall | Single | Inter-pausal Unit | - |
| Levow & Wang (2012) [100] | | ✓ | $F_1$ | Single | Inter-pausal Unit | - |
| Poppe *et al.* (2012) [121] | ✓ | | - | - | Continuous | - |
| Skantze (2012) [133] | | ✓ | Precision | Single | Inter-pausal Unit | - |
| Ozkan & Morency (2013) [117] | | ✓ | $F_1$ / UPA | Single | Continous | 0/1000ms |

**Table 6.2:** Overview of the corpus based listener response prediction models developed so far.

### 6.4.1 Objective Evaluations

In objective evaluations of listener response prediction models the predictions made by the models are compared to the ground truth. A measure is selected which quantifies the comparison. Measures that are used to report objective evaluations include cross-correlation coefficient [154], precision [133], recall [110], precision and recall [115, 114, 153, 136, 124] or $F_1$ (which is the weighted harmonic mean of precision and recall) [28, 88, 107, 108, 116, 118, 117, 100]. Most authors opt for a measure based on precision and/or recall, but in two areas differences between measures remain, namely ground truth selection and margin of error.

#### Ground Truth Selection

The majority of evaluations of listener response prediction models are performed by comparing the predictions made by the model with the listener in the corpus [115, 153, 28, 136, 88, 107, 108, 116, 118, 124, 110, 100, 117]. As Ward and Tsukahara [153] have noted this is not ideal. 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*. [28] dealt with this problem by only using high backchannel rate data as test data in order to minimize false negatives. Skantze [133] had one annotator annotate each inter-pausal unit as either inappropriate, expected or optional. For the experiments only the inter-pausal units annotated as inappropriate or expected were used during training and evaluation.

Others have dealt with this problem by collecting multiple perspectives on appropriate times to provide a listener response. This was either by asking multiple people to press a key on a keyboard at times they would give a backchannel in reaction to a recorded speaker [114, 79, 78] or by asking multiple people to intentionally nod [154].

Recently a measure has been proposed that is specifically aimed at being applied to such multiple perspective data. Ozkan and Morency [117] have proposed User-Adaptive Prediction Accuracy as an evaluation metric for listener response prediction models. For this measure the model is asked for $n$ most likely listener response moments in reaction to a speaker, where $n$ is the number of listener responses given by the ground truth listener. This measure allows evaluation of the ability of the model to adapt to different listeners. Some listeners may give a response frequently, while others give a response only a limited number of times.

#### Margin of Error

For the models evaluated using precision and recall based measures on continous data another discriminating factor applies, namely the margin of error. Precision and recall based measures rely on the evaluation of whether a prediction is 'at the same time' as the ground truth. The definition of 'at the same time' differs between evaluations. Okato *et al*. [115] use a margin of error of -100ms to +300ms from the onset of the ground truth listener response, Ward and Tsukahara [153] use a margin of error

of -500ms to +500ms, Poppe *et al*. [124] use a margin of -200ms to +200ms, and Morency *et al*. [107, 108] and Ozkan *et al*. [116, 118, 117] use a margin of error of 0ms to +1000ms.

### 6.4.2 Subjective Evaluations

When it comes to subjective error measures several strategies have been used to establish the performance of the models. The approaches used so far either evaluate a general impression of the listening behavior or individual listener responses.

Fujie *et al*. [51] made a pair-wise comparison between models in which the general impression of the listening behavior was measured. A subject interacted twice with a conversation robot system whose listening behavior was driven by two different models. After these interactions the subject was asked on a 5 point scale, which system they preferred, with 1 being system A, 5 being system B and 3 being no preference.

Huang *et al*. [79] also evaluated the general impression of the listening behavior. They generated virtual listeners in response to recorded speakers and presented these interactions to 17 subjects. Similar to Fujie *et al*. [51] the subjects were presented with three different virtual listeners each driven by a different listener response prediction model. After each interaction the subject was asked 7 questions about their perceived experience with regard to the timing of listener responses. On a 7-point Likert scale the subjects rated the virtual listeners on 'closeness', 'engrossment', 'rapport', 'attention', 'number of inappropriate listener responses', 'number of missed opportunities' and 'naturalness'.

Poppe *et al*. [124] also let participants evaluate virtual listeners in interaction with recorded speakers. They asked participants for each fragment "How likely do you think it is that the listener's behavior has been performed by a human listener". The participants made their judgement by setting a slider that corresponded to a value between 0 and 100.

Kitaoka *et al*. [88] had 5 subjects rate each generated listener response individually. The data presented to the subjects were 16 to 18 samples of single sentences followed by a backchannel. Each generated listener response rated was rated on a 5-point scale ranging from 'early' to 'late', with an extra option for 'outlier'. They carried out this process for listener responses generated at times predicted by their model and times as found in the corpus. The authors accumulated the counts of the 5 subjects and reported the percentage of ratings in the "good" category (rating 3). The same approach was used by Nishimura *et al*. [113].

In a later experiment Poppe *et al*. [121] evaluated their model in a switching Wizard-of-Oz setup. Two participants had an avatar mediated interaction. The speaker would see an avatar representation of the listener. While holding a monologue about different topics, the speaker had to judge the listening behavior of the avatar on the screen. The listening behavior of the avatar on the screen switched several times during the interaction between behavior generated based on a prediction model and based on the listening behavior of the other participant. The speaker was asked to hit the spacebar each time he thought the listening behavior of the avatar was unnatural.

## 6.5 Conclusion

This survey has shown that a wide variety of listener response prediction models exist and that these models have been evaluated in almost as many different ways. This makes comparing different methods in terms of performance even more complicated than it already is. Most models are trained and tested on different corpora, which differ in language, type of conversations and amount of data. This already makes a comparison between reported values unreliable. On top of these differences the evaluation methods used also differ from each other. So, it is hard to say which model is best, although some papers include comparisons to previous methods.

Most previous models simply combined all data available into one big training set. By doing this individual differences in speaking and listening behaviors are discarded and therefore valuable behavioral patterns from outliers are lost in the process. In the following chapters three new listener prediction models will be introduced that embrace these individual differences and similarities.

The first model, presented in Chapter 7, focuses on using the consensus perspective from the parallel recorded listeners in the MultiLis corpus. The model uses the consensus perspective during learning as well as evaluation in a novel way to make the listener response prediction model sensitive to the graded optionality of the response opportunities in the corpus. The model values the response opportunities with multiple responses more highly, while not ignoring the response opportunities where only one listener responded.

The second model, presented in Chapter 8, focuses on using the perspectives obtained using the individual perceptual evaluation method (Section 3.2) as negative samples during learning to have a more accurate ground truth. Through iterative learning and perceptual evaluation the model is refined and additional training data becomes available.

The final model, presented in Chapter 9, focuses on learning a model for each speaker individually instead of learning a single prediction model for all speakers as is usually done. Speakers can differ wildly in how they behave when speaking. Some may look at the listener often and long, while others only occasionally for a brief period. Some are fluent in their speech, while others often hesitate. If a model is dependent on a certain cue (e.g. the speaker looking at the listener) and a speaker does this less frequently than most of the speakers in the training data, the model will most likely fail to work properly. General models tend to gravitate to modeling the behavior of the majority of the speakers and behavioral patterns from outliers are often lost in the numbers. With a collection of models trained on each speaker indepently, when a new speaker is encountered one can find the speaker from the training data that behaves most similarly and use the model that has been trained for that speaker.

# 7

# Learning and Evaluating Using the Consensus Perspective

In Section 2.3 the consensus perspective was introduced. The consensus perspective was defined as the complete coverage of all identified response opportunities in a corpus and for each identified response opportunity, the number of listeners that responded to that response opportunity. This consensus perspective captures the differences and similarities between individuals with respect to listening behavior. In this chapter we will explore the advantages the consensus perspective brings to the prediction of listener responses. In particular we will look into how this consensus perspective can be used to select more accurate positive samples for the ground truth labels.

How the consensus perspective will be used to improve the state of the art of listener response prediction in both learning and evaluation will be discussed in Sections 7.1 and 7.2 respectively. This will be followed by a description of the experimental setup that was used to evaluate these contributions in Section 7.3. Results of the experiments will be presented and discussed in Section 7.4, which will be followed by the conclusion in Section 7.5.

## 7.1 Using Consensus Perspective 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 7.1). Our machine learning approach trains a sequential probabilistic model from a database of consensus interactions and uses 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.

The MultiLis corpus collected several perspectives on listening behavior. By only

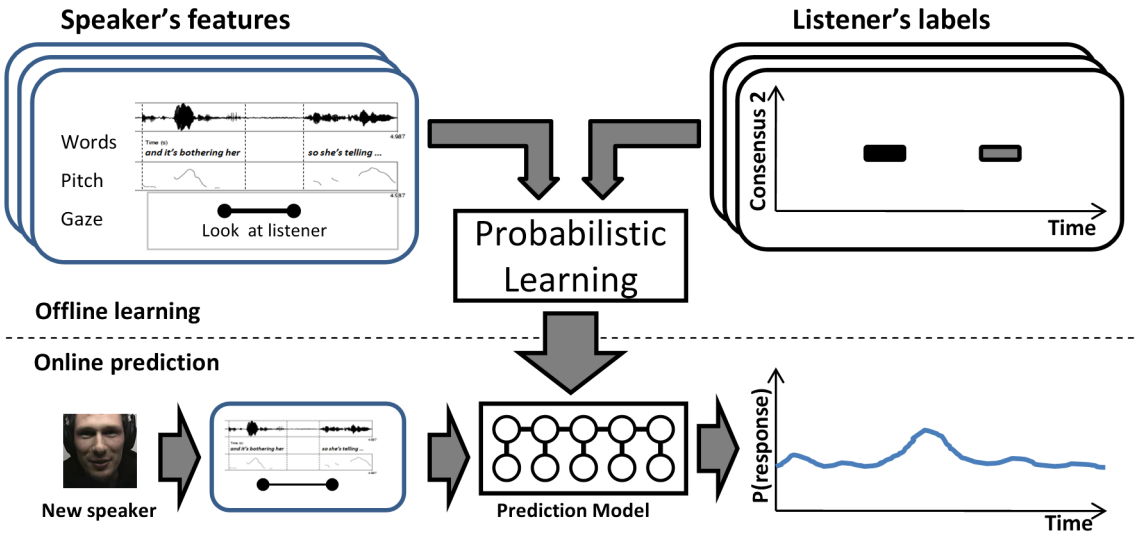


**Figure 7.1:** Learning using parallel listener consensus. A sequential probabilistic model is trained offline using as input a sequence of observation features (e.g., the speaker features) and the ground truth labels from the consensus data (Consensus 2 in this figure). The prediction model returns a sequence of probabilities (i.e., probability of listener response) during the online testing.

considering the perspective from the displayed listener this corpus can be regarded as any other corpus of recorded one-to-one interactions with its limitations caused by individual differences in listening behavior. Some individuals give a lot of listener responses, while others use their responses more sparingly. If 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 the MultiLis corpus also offers the listening behavior of the concealed listeners. These can be used during learning. By recording more people less of our data is mislabelled as negative examples. As was shown in Section 2.3.2 using the displayed listeners gave in total 879 responses to response opportunities, but the two additional perspectives have increased this to identify 1733 response opportunities in the consensus perspective. This is almost a doubling of the number of positive samples.

The consensus perspective 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 responded; whereas to some others, only two or mostly one of the listeners responded. As was shown in Chapter 5, the response opportunities to which three listeners responded are more clearly cued by the speaker than response opportunities to which two listeners responded and even more than opportunities to which only one listener responded. The speaker will have expected a response at these moments. By emphasizing these response opportunities during learning the resulting model should be better tuned to predict these response opportunities and will therefore result in a better model.

## 7.2 Using Consensus Perspective during Evaluation

The consensus perspective can also be used to give a more reliable performance measure during evaluation. We propose a new evaluation criteria which uses the consensus perspective 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. How this is achieved is highlighted by the following two concepts; 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 could have given a response at that moment. With the consensus perspective 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 identified response opportunities as ground truth.

- **Consensus completeness** - If all response opportunities 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 than any individual person. The experiment of Huang *et al*. [79] 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. In this proposition the consensus level means the number of listeners that need to respond to the response opportunity before the response opportunity is included as a ground truth for the computation of recall.

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

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

In Equation 7.2 the precision $Precision_{all}$ is calculated using all response opportunities in the consensus perspective and the recall $Recall_t$ is calculated using only the response opportunities where at least $t$ listeners responded. $t$ 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 the MultiLis corpus, the average response rate is 5.8 responses per minute. The closest match to this is 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. Using this performance measure means that the model will maximize the number of predicted responses at times where at least two listeners responded, while only regarding responses predicted at moments where no listener responded as false positives.

## 7.3 Experimental Setup

In this section the experimental setup will be explained. In this experiment five models were trained that differ in the way the consensus perspective is used to select ground truth labels. The models are evaluated on four performance measures again differing in the way the consensus perspective is used to selected ground truth labels. First, the model and corpus used to train the prediction models will be presented. This will be followed by a description of the features that are used to describe the behavior of the speaker. Finally, the difference between the five models and four evaluation measures will be explained.

### 7.3.1 Machine Learning Model

The machine learning models trained in our experiments were Conditional Random Fields (CRF) [95] and were 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 features describing the behavior of the speaker, and a sequence of ground truth labels, here the onsets of response opportunities from the consensus perspective. The learned model returns a prediction value curve with a value at each frame indicating the probability of response opportunity. After smoothing the prediction value curve can be used to predict response opportunities by detecting peaks in the curve. By comparing the heights of these peaks to a fixed threshold the most probable moments are selected as predicted response opportunities.

We used 31 interactions from our Dutch spoken MultiLis corpus presented in Chapter 2. All models were trained with the same training set of 21 interactions and tested on the same test set of 10 interactions. The test set did not contain individuals from the training set. The objective function of the CRF model contains a regularization term to prevent overfitting. During training, this regularization term

was validated with values $10^k$, for $k = -3..3$ using a 3-fold strategy on the training set.

### 7.3.2 Multimodal Features

The behavior of the speaker is described using acoustic, lexical and visual features.

- **Acoustic Features** - Utterance, pause and pitch are used as acoustic features. Utterence and pause are binary features extracted using the segmentation from the Dutch automatic speech recognition software SHoUT [81]. The minimum length for a pause is 100 ms. Utterences with a pause smaller than 100 ms between them are combined. Pitch (F0) is extracted using openSMILE [46] at a 10 ms interval. To filter out some of the missed frames, the raw pitch values are smoothed with a window size of 5 frames. Finally, the pitch values are discretized into percentiles.

- **Lexical Features** - The words recognized by SHoUT are used as lexical features. Each word is represented as a binary feature which is true at the times the word is spoken.

- **Visual Features** - Eye gaze and blinks are used as visual features. For eye gaze 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 was addressing him/her. Therefore the "continued gaze" feature where the blinks between and after a gaze annotation are included in the interval was created. From both the normal gaze and the continued gaze features a "blinked" variant was created, which only includes the gaze intervals which were preceeded by a blink.

Early fusion is used to combine the features from different modalities. No feature selection is performed. All models are trained using all features.

### 7.3.3 Ground Truth Labels

The goal of the experiment was to find the best way to use our consensus perspective to define the ground truth labels for learning and evaluating listener response prediction models. In this experiment five models were evaluated and compared to each other. Each model differs in the strategy that was used to select the positive ground truth labels.

- **Displayed Listener Only** - Our first model was a CRF model trained only using listener responses of the displayed listener as the ground truth labels. This model was our main baseline for our experiments since most previous work used this approach (such as [28, 108, 153]). This model is referred to as the *DL only* model in the remainder of the chapter.

- **All Listeners** - In the second model, the listener responses of both the displayed listener and the two concealed listeners were used individually as ground truth labels. No consensus perspective was built, but samples where two listeners responded at the same time were simply duplicated in the data set. This model is referred to as the *ALL* model in the remainder of the chapter.

- **Consensus 1, 2 and 3** - The last three models were trained using the consensus perspective. The *Consensus 1* model included all response opportunties from the consensus perspective. So all the 1733 response opportunities to which at least one listener (either the displayed listener or one of the concealed listeners) responded were used as ground truth label. The *Consensus 2* model only included the 593 response opportunities to which at least two listeners have responded as ground truth label and the *Consensus 3* model only used the 128 response opportunities to which all three listeners responded.

In all models the negative ground truth labels were selected at random times where no positive sample was found. The ground truth labels were normalized to the same length of 700ms. For each response opportunity, the mean onset of the responses belonging to this response opportunity is calculated. The ground truth label starts at the 350ms before this mean onset time and ends 350ms after it.

### 7.3.4 Evaluation

The secondary goal of the experiment was to evaluate our proposed $F_{consensus}$ performance measure. Therefore, each model is evaluated on four evaluation metrics that again differed in their ground truth selection.

- $F_1$ **using displayed listener only** - For this measure only the listener responses of the displayed listener were used as ground truth labels. Using these labels the $F_1$ measure was calculated.

- $F_1$ **using consensus 1** - For this measures the consensus perspective was used as ground truth labels. Both precision and recall ere calculated using all response opportunities in the consensus perspective as ground truth labels.

- $F_1$ **using consensus 2** - For this measure both precision and recall were calculated using only the response opportunities with at least two responses as ground truth labels.

- $F_{consensus}$ - For this measure the precision was calculated using all response opportunities as ground truth labels and recall was calculated using only the response opportunities with at least two responses.

## 7.4 Results and Discussion

During our experiments five models were trained using various strategies to establish ground truth. These models were evaluated on four different performance measures. The results of these experiments will be discussed per performance measure.

| 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 7.1:** 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 7.2:** The performance of our five models measured using the Consensus 1 and Consensus 2 ground truth labels.

### 7.4.1 $F_1$ using displayed listener only

In Table 7.1 the performances of the five response prediction models are presented as measure by $F_1$ using only the listener response from the displayed listener as ground truth. This can be seen as the baseline performance of the models as this is the most frequently used way to measure the performance of listener prediction models (see Section 6.4.1). The performance of our baseline *DL only* model ($F_1 = 0.265$) is comparable to the performance of Morency *et al.* [108] ($F_1 = 0.256$). Out of the other models that make use of the additional listeners in the MultiLis corpus the *Consensus 2* model performs best ($F_1 = 0.264$), which is comparable to the *DL only* model. The *ALL* model performs only slightly worse ($F_1 = 0.255$). The other approaches perform less well as the baseline model on this performance measure.

### 7.4.2 $F_1$ using the Consensus Perspective

As discussed in Section 7.3.4 the MultiLis corpus provides us with more information than only the responses of the displayed listener. We also have the responses 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. Because of individual differences 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 responded. A prediction of our model at such a missed response opportunity should not be counted as a wrong prediction, since according to our corpus, these are moments

| 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 7.3:** The performance of our five models measured on our $F_{consensus}$ measure. The difference between our Baseline model and Consensus 2 is marginally significant, $p = 0.054$.

where listeners do provide responses.

The consensus perspective contains all response opportunities to which at least one listener responded. The middle column in Table 7.2 shows the performances of the five models measured by the $F_1$ measure calculated using consensus 1 as ground truth labels. On this measure the *ALL* ($F_1 = 0.377$), *Consensus 2* ($F_1 = 0.375$) and *Consensus 3* ($F_1 = 0.364$) models perform significantly better than the *DL Only* model ($F_1 = 0.278$). The *Consensus 1* model ($F_1 = 0.318$) also performs better than the baseline model, but significantly worse than the other models.

However, this measure does not reflect what people consider to be the behavior of a believable and attentive virtual human. If a response is generated at each response opportunity in the consensus perspective the response rate would be twice as high as the average listener in the corpus. Using a similar data set, Huang *et al.* [79] have shown that a virtual human which responds at moments most people would respond at is the most believable.

In our consensus perspective this corresponds to the response opportunities where at least two listeners have responded. Using these ground truth labels the response rate is closest to the response rate of the average listener. The right column in Table 7.2 shows the performances of the five models measure by the $F_1$ measure calculated using consensus 2 as ground truth labels. Again, the *Consensus 2* model ($F_1 = 0.287$) performs slightly better than the baseline *DL only* ($F_1 = 0.253$), *ALL* ($F_1 = 0.255$) and *Consensus 3* ($F_1 = 0.256$) models.

### 7.4.3 $F_{Consensus}$

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 Table 7.3. The *Consensus 2* model ($F_{consensus} = 0.439$) performs better than the baseline *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 a model using the response opportunities where at least two listeners responded as ground truth performs best in all cases. The proves the useful-

ness of the consensus perspective in the learning phase. Furthermore using $F_{consensus}$ as performance measure gives us a more reliable performance measure which takes advantage of the consensus perspective to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models.

## 7.5 Conclusion

In this chapter, the usefulness of the consensus perspective was shown for both learning and evaluating probabilistic prediction models of listener responses. To improve the learning performance, the consensus perspective helps us to reduce outliers in the positive samples. Across all performance measures learning a model using response opportunities where at least two listeners responded as ground truth labels performed best.

Furthermore, a new performance measure called $F_{consensus}$ was proposed which takes advantage of the consensus perspective to better define the concepts of exactness (mislabels) and completeness (missed labels) for prediction models. By calculating precision using all response opportunities and recall using only the response opportunities where the majority of the listeners responded the performance measured matches the perception of the behavior more closely.

At this time only three parallel recorded listeners were used, but getting more listeners in parallel or by using the parasocial perspectives the improvements these techniques bring could be even greater. More listeners would mean having an even bigger coverage of the response opportunities and therefore fewer false negative samples. Furthermore, a more accurate threshold could be selected for the minimum number of listeners that need to respond to a response opportunity for the ground truth labels (for both ground truth labels in learning and in the $F_{consensus}$ measure), reducing outliers.

# 8

# Learning using Individual Perceptual Evaluation

The previous chapter has shown the merits of the increase in coverage of all possible response opportunities in the MultiLis corpus. This increase in coverage was obtained by the combination of multiple perspectives on listening behavior from the recordings of the three listeners. Beside the collection of perspectives on appropriate timing of listener responses Chapter 3 also introduced the collection of multiple perspectives on *in*appropriate timings of listener responses. These perspectives were collected using the Individual Perceptual Evaluation method presented in Section 3.2. This chapter will show how the perspectives obtained using this method can be used in the development of listener response prediction models. The main focus will be on selecting better negative samples for the ground truth labels.

## 8.1 Iterative Perceptual Learning

The ground truth labels that are used in the learning and evaluation of listener response prediction models consist of examples of both positive and negative moments of the behavior. Most attention is usually paid to the collection of positive samples, moments where a listener response is appropriate. This is done by recording a corpus in which the listener responses that are given by the interlocutors are carefully annotated. In contrast, negative samples are usually simply collected by making a random selection of moments that do not overlap with the positive samples. However, as was argued before, it is not certain that a moment where no listener response is given is actually an inappropriate time for such behavior. Giving a listener response is optional. It is likely that one or more of those negative samples are actually mislabelled positive samples. These mislabeled samples hinder the learning of the prediction models. By using the consensus perspective in our previous model the chances of this happening were reduced, but not eliminated. Only three perspectives were combined and the parasocial perspectives collected in Section 3.1 already showed that additional perspectives still identify many other response opportunities.

With Individual Perceptual Evaluation a method was introduced that can produce the ground truth negative samples needed for learning our prediction model. In this

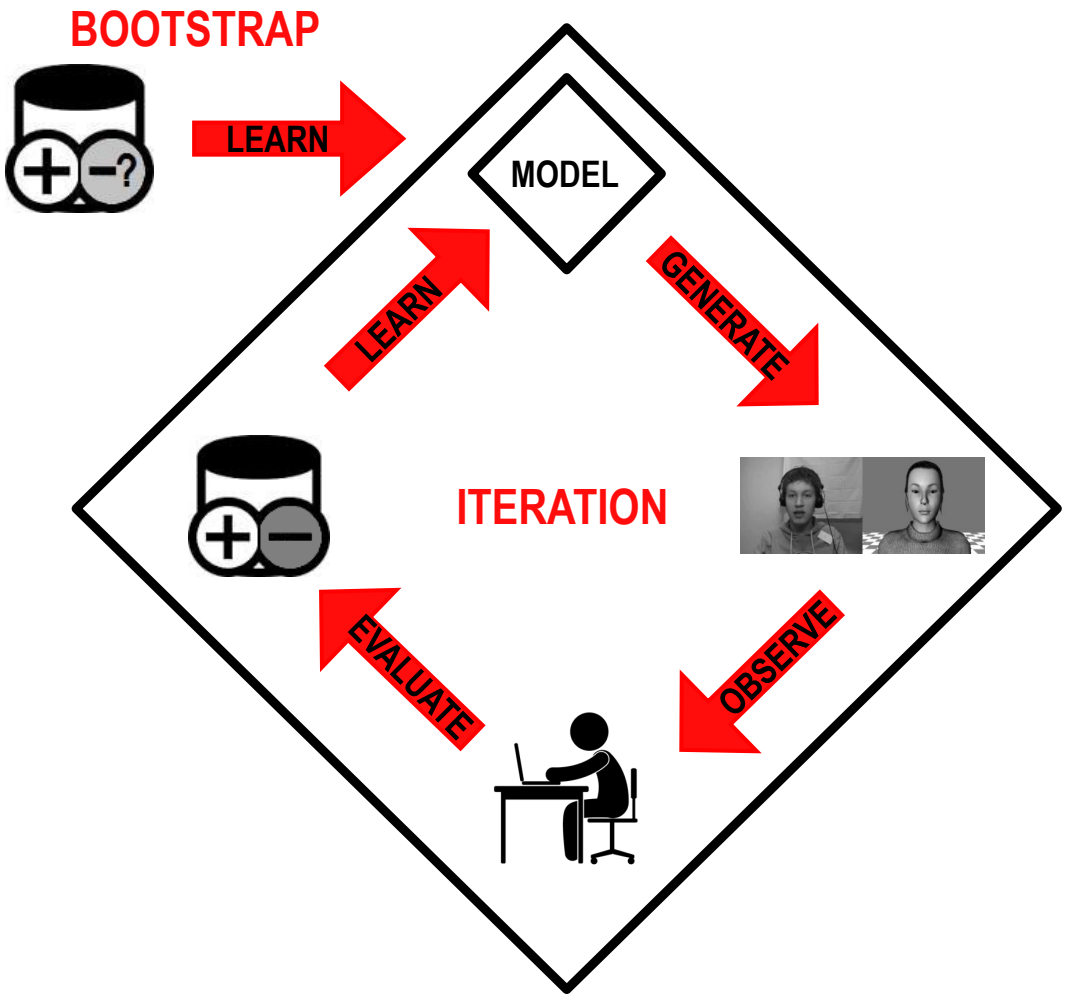


**Figure 8.1:** Overview of the proposed iterative perceptual learning approach. Initially in the bootstrap phase a model is learned on a set of interactions with positive labels and randomly selected negative labels. This model is used in the first iteration to generate listening behavior on new interactions that a human judge evaluates. The subjective ratings resulting from this evaluation are used in following iterations as verified negative labels.

method humans observe a virtual listener and rate the timing of the generated listener responses. Each time the humans observe an inappropriate listener response they hit a key on the keyboard. Beside an evaluation of the performance of the model that generated these listener responses, these subjective ratings can act as negative training samples for subsequent models. In this chapter an approach is proposed that uses such ratings as negative samples. This approach is called the iterative perceptual learning approach.

In Figure 8.1 the development cycle of the proposed iterative perceptual learning approach is illustrated. The development cycle starts with learning a bootstrap model. This bootstrap model is trained on a (small) subset of the corpus. The positive samples of the ground truth are the listener responses found in the corpus and the negative samples are randomly selected. This bootstrap model is used to generate listening behavior in response to the speaker of a new set of interactions. Next, the

generated listening behavior is observed and judged by human observers using the individual perceptual evaluation method. Then, a new model is learned using the positive samples from the listening behavior of the original listener in the corpus and the negative samples from the subjective ratings of the generated listening behavior. Using the same methodology this new model can then be evaluated again.

Besides the fact that using the subjective ratings as negative samples gives us verified ground truth negative samples, there is another benefit to this approach. The negative samples are also at moments where the current model makes mistakes. This means that if the process of learning and evaluating is repeated multiple times on new data, the positive and negative samples should be selected at more relevant moments each iteration. If the expressiveness of the features and the modelling power of the machine learning model are strong enough, earlier mistakes should be avoided after each iteration.

This approach is a variant of active learning. The type of learning has proven to be a succesful method in improving training results in areas such as speech recognition, information extraction and media classification [131]. Here, the application of the technique on social behavior, in particular listening behavior is in question.

## 8.2 Experimental Setup

To test the Iterative Perceptual Learning (IPL) approach an experiment was set up that compared the IPL approach to a baseline model. The IPL model was learned in one bootstrap phase and four iterations. After each iteration the subjective ratings were used to learn the new version of the IPL model. At each stage of the development a baseline model was learned on the same data, but with random negative samples instead of subjective ratings as negative samples. The models were evaluated using these subjective ratings as well as objective measure comparing their predictions to the listener responses found in the corpus.

In the remainder of this section more details of the experiment will be presented. First the machine learning techniques will be explained, followed by a description of the features. Then the ground truth labels will be presented and finally the evaluation method.

### 8.2.1 Machine Learning Model

The machine learning models trained in our experiment were Support Vector Machines (SVM). Our interest lays in the relative performance of both approaches and did not focus on obtaining an optimally performing model. Therefore, the default settings of the libSVM library [31] were used without optimization of the parameters involved. These settings were an RBF kernel with $c = 1$ and $\gamma = 1/|x|$, where $|x|$ is the dimensionality of the input vector.

For training a feature vector was extracted at each positive or negative sample according to the ground truth labels. After training the SVM model was applied to feature vectors taken at each frame on a 10 millisecond interval. To obtain the listener response prediction the numerical decision values were used. These can be

regarded as confidence scores. By sequencing these decision values over time, a prediction value curve is obtained representing the appropriateness to provide a listener response. To remove artifacts due to the potentially highly non-linear output of the SVM, this curve was smoothed with a 10 frame moving average. After this filtering, the highest peaks in this curve were regarded to be the most likely moments to predict a listener response. A fixed threshold was used to determine at which peaks a listener response should be generated.

### 8.2.2 Multimodal Features

As mentioned earlier the SVM generates predictions in response to a feature vector. This feature vector describes the behavior of the speaker. For this experiment 113 acoustic and 1 visual features were extracted from the audio and the video signal.

- **Acoustic Features** - As acoustic features pitch, intensity and the first 12 mel-frequency ceptrum coefficients (MFCC) were extracted from the audio signal using OpenEAR [46] at a frequeny of 100 Hz. Pitch detection is typically noisy and can fail for a few frames during speech. To solve this issue, gaps in the pitch values smaller than 8 frames were linearly interpolated, which is in line with [153]. Between subjects, acoustic signals can vary significantly. For instance, pitch is higher in females than in males, people speak with different volume and/or had the microphone closer to their mouth. To account for these differences between speakers the features were normalized by converting each signal into the $z$-score equivalent. The means and standard deviations needed for calculating the $z$-score were obtained from the first 10 seconds of each session, which were excluded from the training data.

  To aid the support vector machines in capturing the temporal relation between the acoustic cues and the onset of a listener response, some of the temporal aspect was encoded into the features. To this end the mean and the slope of each signal were calculated over a period of 50ms, 100ms, 200ms and 500ms prior to the onset of a listener response. As such, the behavior that was expected to cue the listener responses was captured. The slope was calculated by fitting a first-order polynomial to the signal.

  Finally the speech/non-speech features were extracted using the SHoUT automatic speech recognizer [81], indicating whether the speaker is talking or not.

- **Visual Feature** - Eye gaze was used as a visual feature. The feature is based on the manual annotations of whether the speaker was looking at the listener or not.

Both the speech/non-speech and eye gaze features were initially binary. To represent sequentiality, the relative offset to the moment where the speaker started talking or started looking at the listener was calculated. Specifically, positive numbers denote the offset from the start of the speaker's speech and negative numbers the offset from the start of a pause.

In summary, 14 acoustic signals were extracted, their $z$-scores were calculated and their means and slopes for four different window lengths were obtained. A

speech/non-speech feature was added for a total of 113 acoustic features. Eye gaze was added as a final visual feature. Following the early fusion paradigm, all features were concatenated into a 114-dimensional feature vector per frame. During learning no feature selection was performed to selected the optimum feature set. All models were trained using all features.

### 8.2.3 Ground Truth Labels

In Chapter 7 the difference in ground truth labels was in the positive samples. In this Chapter the models that are compared differ in the selection of negative samples for the ground truth labels.

- **Baseline Model** - In the baseline model the negative samples were obtained from random selection. The samples were taken from the moments where no listener response was given by the displayed listener. Due to the optionality of backchannels, they possibly included false negatives. Typically, there is only a small number of positive samples available in a corpus.

- **Iterative Perceptual Learning Model** - For the iterative perceptual learning model (IPL model) the negative samples were obtained from the judgments collected in the evaluation of each iteration. After a bootstrap model was trained on one interaction using the baseline model approach, a virtual listener was generated based on this model. The behavior of this virtual listener was evaluated using the individual perceptual evaluation method presented in Section 3.2. The timing of generated listener responses that were judged by the participants of the evaluation were used in the following iteration as negative samples in the training phase. This process was repeated for four iterations, each with an increased number of interactions as training and evaluation data.

Positive samples for both models are the moments where the displayed listener provided a listener response. To increase the amount of training data and to make the models less dependent on single frames, four additional frames around the positive frame were selected using a normalized Gaussian distribution with a $\sigma$ such that 95% of the samples fell within 250ms of the positive sample.

For both models it was ensured that there was an equal number of positive and negative samples. For the baseline model this was done by selecting the same number of negative samples as there were positive samples in the ground truth. For the IPL model the weight of each yuck response was determined by the ratio between the number of positive samples and yuck responses. This ratio determined the number of additional frames that were selected within 250ms of the negative sample. Listener responses that received a multiple number of yuck responses were added as many times as they were yucked.

### 8.2.4 Evaluation

Both models were evaluated on the corpus objectively and in an individual perceptual evaluation experiment. The individual perceptual evaluaton experiment was per-

| Phase | Training Set | # of interactions | Evaluation Set | # of interactions |
|---|---|---|---|---|
| Bootstrap | Bootstrap set | 1 | Set 1 | 1 |
| Iteration 1 | Set 1 | 1 | Set 2 | 2 |
| Iteration 2 | Sets 1, 2 | 3 | Set 3 | 3 |
| Iteration 3 | Sets 1, 2, 3 | 6 | Set 4 | 6 |
| Iteration 4 | Sets 1, 2, 3, 4 | 12 | Test set | 6 |

**Table 8.1:** Overview of the sets used in each iteration and each phase of the IPL process.

formed in one bootstrap phase, followed by four iterations. In each phase the data set on which the models were trained and evaluated was increased.

### Procedure

The experiment consisted of five phases. It started with a bootstrap phase, followed by four iterations in which the IPL model was built. In the bootstrap phase, a baseline model was learned on a single interaction. This model was then evaluated perceptually on one other interaction to obtain the first set of negative samples.

In the following four iterations the positive and negative samples obtained from the previous evaluation sets were used as training set for the IPL model. Given that the negative samples were selected at random in the bootstrap phase, all samples of this phase were discarded for the first iteration. Since the evaluation results for the IPL model doubled as negative samples for model learning, there were more positive and negative samples available to learn the IPL model in each subsequent iteration. In addition, a larger set of interactions was used for each iteration. An overview of the number of interactions used for learning and evaluation is given in Table 8.1.

To compare the performance of the IPL model a baseline model was trained for each iteration as well. This model was trained on the same interactions according to Table 8.1, but with negative samples selected randomly without overlapping with positive samples. A fair comparison can be made, since both models were trained on the same interactions and rated by the same participants. Beside the evaluation sets, the models were also evaluated on a test set of six interactions that was never used for training.

Participants of the experiment were shown stimuli through a webpage. It was explained to them that they would be participating in an experiment to determine the quality of synthesized listening behavior. After entering their name, gender and age, the participants were presented a set of (at most) 6 stimuli. They were asked to press the spacebar each time the virtual listener performed a backchannel they judged as inappropriate (a yuck response). Participants could replay the stimulus from the start, which would discard all previously issued yuck responses for that stimulus. Each participant was shown the same interaction twice: once with the virtual listener based on IPL, once based on the baseline model. The order of the stimuli was varied systematically. The within-subject design allowed us to evaluate the difference between the two models pair-wise. This is essential as there are typically differences in the number of yucks between participants. An experiment session lasted around 30 minutes.

**Figure 8.2:** Screenshot of one of the stimuli used in the perception study. The participants were presented with a speaker (on the left) and a generated listener (on the right).

### Stimuli

Participants of the experiment were shown a video of a speaker from the MultiLis corpus side-by-side with an animated listener, see Figure 8.2. The virtual listener nodded her head when the synthesis model predicted a listener response. Other behaviors such as head movement, posture shifts, facial expressions and eye blinks were not animated to prevent these factors from contributing to the perception. As a result, the synthesized listening behavior was completely controlled, but rather minimal. For each interaction in a set we created an animation of the virtual listener based on the IPL model and a virtual listener based on the baseline model. The mean duration of a stimulus video was approximately four minutes, depending on the interaction between the actual speaker and listener in the corpus.

Each stimulus video was generated using the Elckerlyc platform [142] to synthesize head nods as listener responses for the listener at the timings predicted by the trained SVM models. To control for the number of listener responses, the mean listener response rate over all interactions in the corpus was used, which is approximately 7.7 per minute. To collect more negative samples to be used in subsequent iterations the response rate was increased an extra 25% to 9.6 listener responses per minute. For both the baseline and IPL models the fixed threshold for the peak selection was such that the listener response rate was matched in each stimulus video. The only restriction applied was that two listener responses could not be within two seconds of each other.

### Participants

Each stimulus was rated by five participants. To limit the time commitment of the task, set four and the test set were split in two, since these sets contained six interactions (12 stimuli). Including the evaluation on the test set for iterations 1 to 4, this gave us 13 groups of stimuli. In total 65 participants were required to rate the stimuli, 25 for the evaluation of sets 1 to 4 and 40 for the evaluation of the test sets. Participants were recruited among colleagues and students. Several persons participated more than once, each time on a different set. As a within-subjects design was used, this does not bias the comparison between the models. Of the 65 trials, 8 were completed

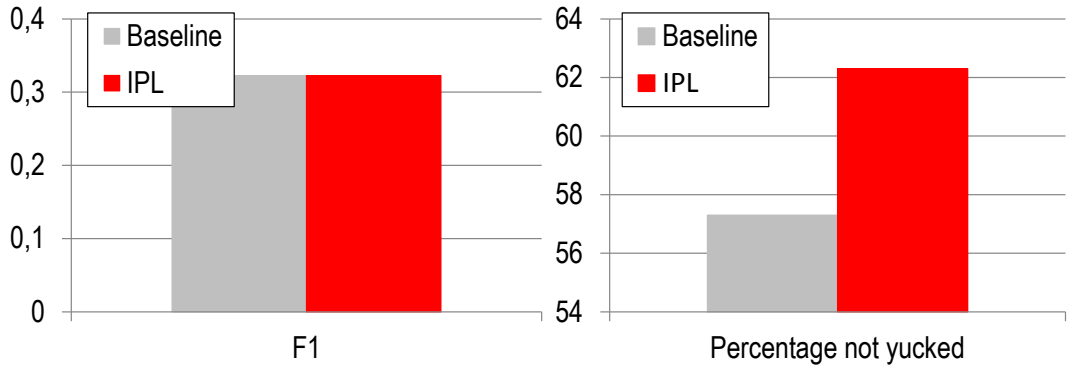


**Figure 8.3:** The performance of both the baseline and IPL model on the objective measure (left) and subjective measure (right). It shows that on the objective measure the models perform the same, but on the subjective measure there is a preference for the IPL model.

by females and 57 by males (mean age 28, min. 18, max. 47).

### Evaluation Measures

The performance of the baseline and IPL models were compared on both objective and subjective measures.

- **Objective measure** - For the objective measure, the predicted timing of the listener responses were compared with the listener responses performed by the displayed listener in the MultiLis corpus. A listener response was regarded as correctly predicted if it was timed within 500ms (before or after) a listener response produced in the corpus. The precision $p$ and recall $r$ were calculated. Precision is the number of matches amongst all predicted signals, recall is the amount of matches amongst all relevant instances in the corpus. Precision and recall were combined in the $F_1$ measure by taking the weighted harmonic mean of the two.
- **Subjective measure** - For the subjective measure the yuck responses collected in the perceptual evaluation were used. The percentage of listener responses that did not receive any yucks was calculated, as well as the average number of yuck responses per listener response.

Performance on these measures was evaluated at each iteration on both the current evaluation set of the iterations and the test set.

## 8.3 Results and Discussion

The performance of both models on the test set after the final iteration is presented. In the left graph of Figure 8.3 the performance of both models on the objective $F_1$ measure is presented. Both models perform the same with $F_1$ scores of 0.323.

The subjective measures show a slightly different effect. In total, 239 listener responses were generated with each of the models. The number of yuck responses

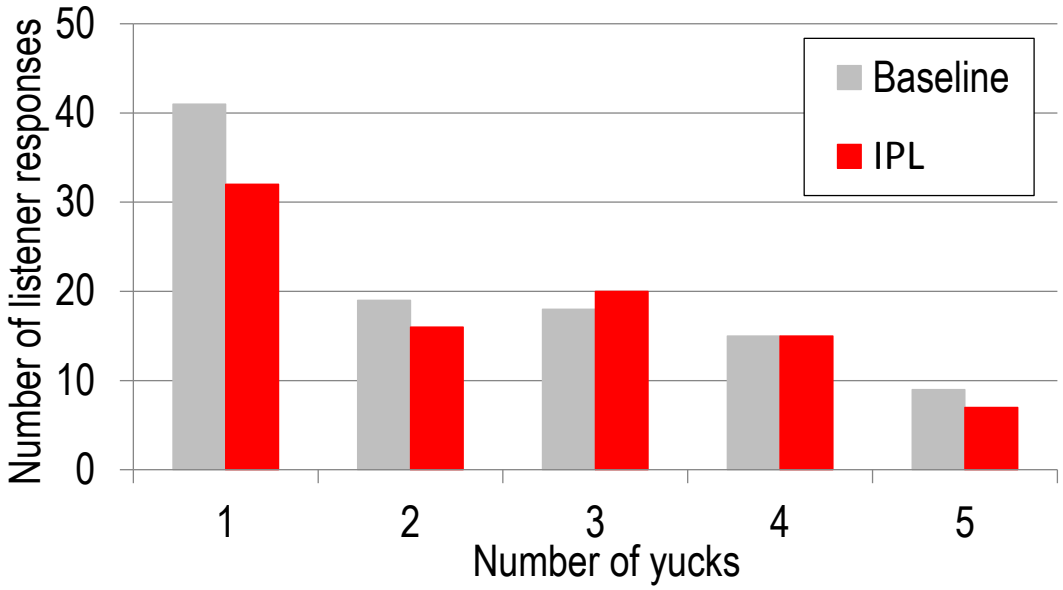


**Figure 8.4:** Frequency histogram for number of yuck responses per synthesized listener responses on the test set, for baseline and IPL models after iteration four.

obtained from five participants per stimulus is lower for IPL than for the baseline (219 and 238, respectively). A pair-wise t-test for the number of yuck responses per stimulus shows a marginally significant difference between the two models ($t(5) = -1.516$, $p = 0.09$). On average, a listener response synthesized with IPL received 0.92 yuck responses from five participants, whereas this number was 1.0 for a listener response generated from the baseline model. A breakdown of the number of yucks per listener response is given in Figure 8.4. Most of the generated listener responses received a modest number of yuck responses.

The number of listener responses that did not receive any yucks is higher for the IPL model, 149 (62.3%) compared to 137 (57.3%) for the baseline model (see also Figure 8.3). In conclusion, both models generate behavior that approximates that of the listener in the corpus in terms of co-occurring listener responses, but the behavior generated based on the IPL model is perceived as slightly more natural.

## 8.4 Conclusion

In this chapter the Iterative Perceptual Learning approach was presented. IPL takes an active learning approach to develop the listener response prediction model. The approach uses the individual perceptual evaluation method to collect subjective ratings from human observers of the predictions made by the model. By iteratively using these subjective ratings as negative samples for the next model, the models become more attuned to the behavior that is perceived as appropriate by human observers.

The merits of the approach were demonstrated by comparing the performance of the IPL model to a baseline model where negative samples are randomly selected at moments that do not overlap with the positive samples. In this comparison both

models performed the same when compared to the original listening behavior found in the MultiLis corpus, but the IPL model received fewer yuck responses in total and the number of individual responses that were found to be appropriate by all five human observers was also in favor of the IPL model.

# 9

# Speaker-Adaptive Learning

Having a conversation requires complex coordination between the interlocutors. This coordination occurs in all modalities of behavior. The content of the speech is chosen such that a coherent conversation unfolds. During the conversation the nonverbal behaviors are also utilized to regulate turn-taking, emphasize important parts of the interaction and establish rapport with the interlocutors. It is a constant back and forth where actions are chosen depending on the actions taken by the other interlocutor(s). The adaptation towards the other interlocutor(s) shows in their speech through changing voice levels, utterance frequency and pauses [64], as well as visual behaviors such as postures, facial expressions and other gestures [32].

This collaborative coordination is not only limited to speaking behavior, but can be found in listening behavior as well [7]. Listener responses are placed at specific places in the discourse. Oftentimes the speaker cues these places and expects a listener to respond [63]. The absence of the expected listening behavior at such places can result in restarts (and often rephrases) from the speaker [56]. This affects the fluency of the conversation, which in turn affects speaker clarity and ultimately speaker comprehension [93, 7]. It has also been proven to hurt the rapport between interlocutors [62].

So far, researchers have focussed on building a prediction model that generalizes all speakers and listeners into one model. The thought process behind this development strategy is that the machine learning techniques pick up on the similarities between the speakers and produce a model that works well in most cases. By doing this individuality in behavior is ignored. The previous models in this thesis have focussed on making the generalization models as inclusive as possible by collecting multiple perspectives on appropriate and inappropriate listening behavior.

In this chapter the focus is on developing a model that explicitly models the differences in behavior. Specifically it adapts to the speaking style of the interlocutor the model is responding to. The approach aims to solve the following issue with the generalization models. An often used feature in listener response prediction models is eye gaze. However, not every person is as comfortable with looking other people in the eye during conversations as others and will do this less often. Therefore, the

model will probably not perform as well for this speaker, when a prediction model is heavily dependent on this cue. If the cue never comes the model will not predict a listener response as often as desired.

In this chapter a speaker-adaptive listener response prediction model will be presented that does not ignore the individuality of the interlocutors. It adapts to the behavior of the speaker. 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. This representation will be called ”speaker descriptors” and it will be a central component used to match new speakers with their closest prototype.

An extensive set of experiments on the MultiLis corpus will be presented and a comparison will be made between our approach and previously published models on the same dataset. Besides the merits of speaker-adaptation our experiment highlights the importance of using multimodal speaker descriptors when comparing speakers to select the right model from the collection.

The chapter will continue in Section 9.1 with a more detail description of the approach to the speaker-adaptive listener response prediction model. The experiment to evaluate the proposed model will be presented in Section 9.2. The results of this experiment will be presented and discussed in Section 9.3. The chapter will be concluded future directions for the work will be presented in Section 9.4.

## 9.1 Speaker-Adaptive Learning

The core of our speaker-adaptive prediction model is the model collection. The model collection consists of a collection of listener response prediction models that represent the variability in speaking styles found in the corpus the model is learned on. How the prediction models in the model collection represents the variability in speaking style will be described in more detail in section 9.1.2. The model collection is built during an offline learning phase and used in the online prediction phase to adapt to the speaker.

In Figure 9.1 the two phases of developing and using our speaker-adaptive model are illustrated. In the top half of the figure the offline learning phase is depicted. From each speaker and listener pair in the corpus a prediction model is learned. Each 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. Each prediction model is accompanied by the speaker descriptors that describe the overall behavior of the speaker the model is learned on. Section 9.1.1 will discuss this issue.

In the bottom half of the figure the online prediction is depicted. The models that are learned in the offline phase and the associated speaker descriptors are collected in the model collection. When the speaker-adaptive prediction model encounters a

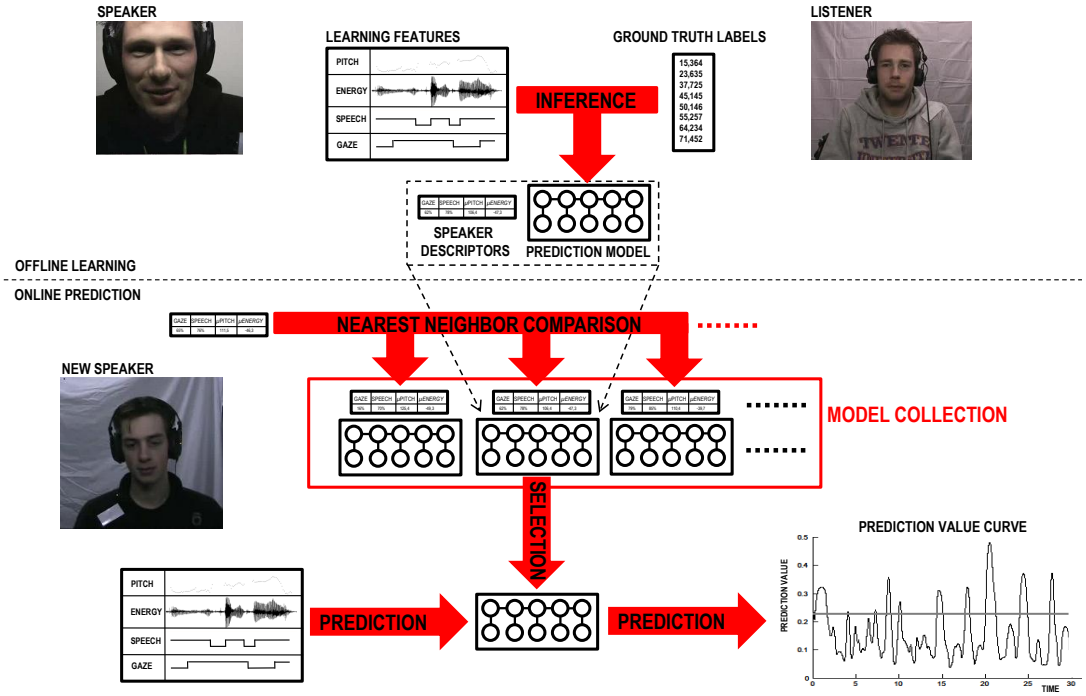


**Figure 9.1:** The figure illustrates the learning and usage of our speaker-adaptive prediction model. In the top half the offline learning phase is depicted. In this phase the prediction models are inferred from individual speaker-listener pairs and each model is placed in the model collection with their associated speaker descriptors. During the online prediction phase (bottom half of the figure) the speaker descriptors from the new speaker are compared to all speaker descriptors in the model collection. The model of the nearest neighbor match is selected to predict listener responses for the new speaker.

new speaker, it calculates the speaker descriptors for this speaker. These new speaker descriptors are compared to all speaker descriptors in the model collection. When the nearest neighbor is found, based on the speaker descriptors, the associated prediction model is selected to predict the listener responses for the new speaker. 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.

### 9.1.1 Speaker Descriptors

One of the novel challenges the speaker-adaptive listener response prediction model introduces is the challenge of finding 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 [139]. 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.

In conversation analysis literature features that are often found to be associated with listener response opportunities include the pitch [89, 153, 63] and energy [89, 63] of the speech signal, pauses in speech [40, 28] and the eye gaze of the speaker [86, 6, 8]. Therefore, it is to be expected that differences lie in these same features. Thus, our focus for the speaker descriptors was directed towards these features.

Each 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 (for instance of pitch and energy) the speaker descriptors that are used are the mean and standard deviation of the signal. For binary features (for instance speech segments and eye gaze) the speaker descriptors that are used are percentage of true values and number of segments per minute.

To select a prediction model the speaker descriptors are compared to all speaker descriptors in the model collection. There are many ways to compare two vectors and find the closest match. We chose nearest-neighbor measured by Euclidean distance for our model. The results presented in Section 9.3 will show that speaker adaptation on these basic speaker descriptors and straightforward nearest neighbor selection will improve prediction performances compared to the a state-of-the-art CRF model without speaker adaptation.

### 9.1.2 Model Collection Composition

As stated before the speaker descriptors 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 balancing 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.

## 9.2 Experimental Setup

In this section we will report the experiment that has been conducted. The goals of our experiments were to (1) compare our speaker-adaptive approach with a priori state-of-the-art approaches, and (2) study the effect of modalities with our speaker descriptors.

The section will be started 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.

### 9.2.1 Corpus

Again the MultiLis corpus presented in Chapter 2.1 was used. However, instead of using all listeners only the displayed listener was used. This decision was made to prove that the approach can work on any corpus and not only on a corpus with the unique characteristics of the MultiLis corpus. The total number of listener responses that were used as ground truth labels in the experiments was 886.

### 9.2.2 Model Learning

The machine learning models trained in our experiments were Conditional Random Fields (CRF) [95] and they were 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 number 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 were compared, the baseline model and a model using the technique explained in Section 9.1. For this technique several individual models were learned. The baseline and individual models were learned as follows:

- **State-of-the-art CRF Model** - Thirty-two state-of-the-art 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. This model was similar to the model of Morency *et al.* [108].
- **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 9.2.4).

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

- **Pitch** - The raw pitch values were extracted using the algorithm of Drugman and Alwan [42] at a sampling rate of 100 Hz. Gaps in detected pitch smaller than 80 ms (8 frames) were linearly interpolated, following [153]. Then all

pitch values were converted to their z-score equivalent. Afterwards the feature was downsampled to 25Hz.

- **Pitch Slope** - As a measurement of the change of the pitch the slope of the pitch value feature was calculated by taking the first derivative of the pitch signal.
- **Energy** - The energy of each speech frame was 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 was calculated by taking the first derivative of the energy signal.
- **Speech Segment** - The speech segment feature captured whether the speaker was speaking at the moment or not. It was represented as a binary feature. The feature was extracted using the segmentation from the Dutch automatic speech recognizer SHoUT [81]. The minimum pause between speech segments was 100ms (4 frames), otherwise speech segments were concatenated
- **Gaze** - The gaze was represented as a binary feature that was true when the speaker looked directly at the listener. The feature was extracted from the annotations provided in the MultiLis corpus.

### 9.2.3 Speaker Descriptors

Our speaker-adaptive model has a model collection. This model collection includes the models that are learned from single speaker-listener pairs. A description of the behavior of the speaker of the pair the model is inferred from is associated with each model. The behavior is captured in 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 presented with a new speaker our speaker-adaptive model calculates the speaker descriptors for the new speaker and compares them to the speaker descriptors found in the model collection. It selects the model whose speaker descriptors are the nearest neighbor match as measured by the Euclidean distance.

### 9.2.4 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 descriptors.

### 9.2.5 Evaluation

The models are evaluated by comparing the prediction 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, for example [108, 117], 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, our learning set is very limited for the individual models 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. See Chapter 10 for more details on this and an alternative solution.

Again, 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.

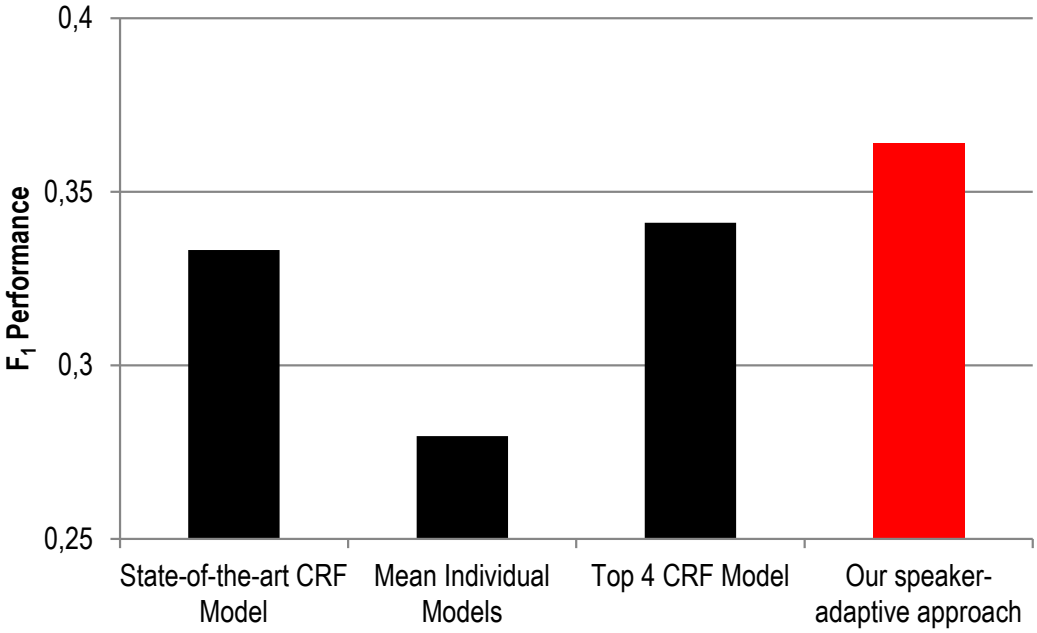


**Figure 9.2:** The figure illustrates the performance of the models included in the experiment. The model proposed in this chapter is represented in red and the models it is compared to in black. The figure illustrates that our speaker-adaptive model performs best with a performance of 0,364. The difference between our speaker-adaptive model and the state-of-the-art CRF model is significant ($t(31) = 3.25, p = 0.001$).

## 9.3 Results and Discussion

In this section the results of the experiments will be presented. The section will start 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 9.3.1. This will be followed by an analysis of the importance of the model collection composition 9.3.2. Finally, the importance of multimodality of the speaker descriptors will be analyzed in Section 9.3.3.

### 9.3.1 Speaker-Adaptation

The performances of the models in question are presented in Figure 9.2. In this figure the performances are represented in red for the models proposed in this chapter and in black for the comparison models.

The performance of our speaker-adaptive listener responses prediction model is an $F_1$ score of 0.364 (fourth bar in Figure 9.2). This is better than 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 9.2). 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 an $F_1$ score of 0.280 (second bar

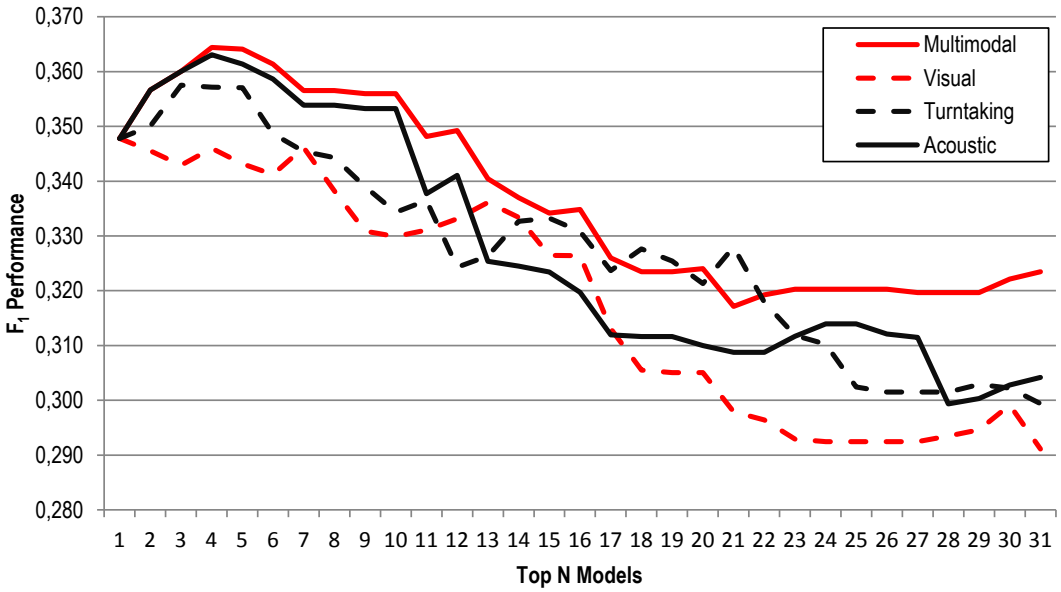


**Figure 9.3:** The figure illustrates two points. First, the maximum performance is achieved by including the top 4 performing models in the model collection with a performance of 0.364. Second, the figure illustrates the importance of the multimodality in comparing speakers. The multimodal nearest neighbor selection (solid red line) almost always outperforms the unimodal nearest neighbor selection.

in Figure 9.2). The best individual model performs at an $F_1$ 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 9.3.2). 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 an $F_1$ score of 0.341 (third bar in Figure 9.2). 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.

### 9.3.2 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. In this section the importance of the composition of the model will be 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 9.3 the results of varying the number of models in the model collection is presented by the solid red line. The other lines will be discussed in Section 9.3.3.

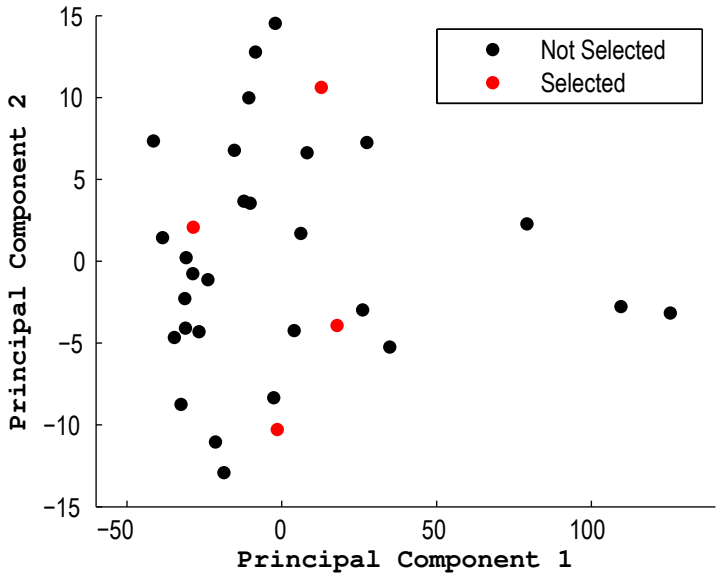


**Figure 9.4:** The figure places each speaker in the 2D space created by the first two principal components of the speaker descriptors. The figure illustrates that the four models that are selected for the model collection (red) are good representatives of the diversity found in the speakers, since they are spread out over the 2D space.

The figure illustrates the maximum performance achieved when the top 4 models are included in the model collection. At this number of models the performance peaks at 0.364 (right bar in Figure 9.2). 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 descriptors space. To analyze this a principal component analysis was made on the speaker descriptors. 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 are presented in Figure 9.4.

In this figure the four speakers that were 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.

### 9.3.3 Multimodal Speaker Descriptors

Finally, an analysis was made of the importance of the multimodality of our speaker descriptors. A comparison was made between multimodal speaker descriptors and unimodal speaker descriptors of the three modalities (acoustic, visual and turn-taking). The comparisons were made on the speaker-adaptive models with varying model collection compositions developed for the previous analysis in Section 9.3.2. The results are presented in Figure 9.3.

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

## 9.4 Conclusion

In this chapter a speaker-adaptive model for predicting listener responses was 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 found in speaker styles found in the corpus as measured by the speaker descriptors. When encountering a new speaker the model compares the speaker descriptors of this speaker to all the speaker descriptors 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 9.3 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). Experiments showed that the speaker-adaptation, the composition of the model collection and the multimodality of the speaker descriptors 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 in 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.

# 10

# Interpreting the Prediction Value Curve

The goal for developing listener response prediction models is applying them in embodied conversational agents to generate appropriate listening behavior. However, during the training and testing these models are often optimized to perform another task, namely matching the ground truth labels found in the corpus as closely as possible. The assumption here is that if we can match the ground truth labels as closely as possible the behavior of the embodied conversational agent will be as good as it can be. But for integrating such models into embodied conversational agents other factors play a role as well.

When designing a conversational embodied agent the designers usually have a personality or role in mind they want their agent to fulfill. A lot of factors are important when managing the impression a user has about the personality of an agent, one of which is their listening behavior. The timing, amount and form of the listener responses that are produced by an embodied conversational agent have been proven to influence the impression the user has about their personality [38].

Therefore, it is important that the produced listening behavior is consistent with the targeted personality of your embodied conversational agent and is so under every circumstance and for every user. In other words, it is important that the listening behavior that an embodied conversational agent performs is stable, recognizable and conform the expectations the user and designer have of the behavior.

This is typically not provided by the current state-of-the-art models for generating generic listener responses. Changes in conditions, such as different interlocutors with different voice characteristics and speaking styles, can have a big impact on the features used as input by these models, which in turn can have a big impact on the predictions made by the models. In this chapter we will propose a technique to make the transition from a corpus based model to a model that is useful for integration into an embodied conversational agent. The technique will focus on a novel way to interpret the prediction value curve that is the output of the listener response prediction models.

All listener response prediction models trained in the previous chapters produce a prediction value indicating the likelihood of a listener response occuring at each time

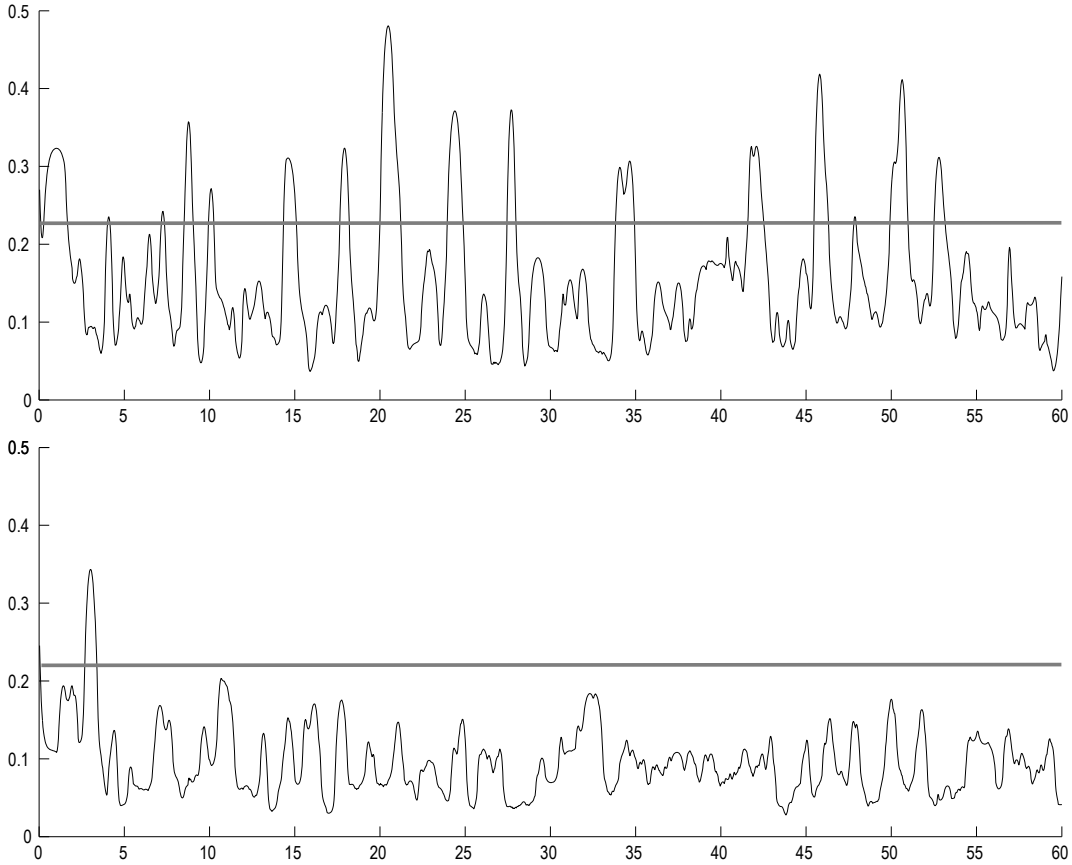


**Figure 10.1:** The prediction value curves of the parallel listener consensus model applied to the first minute of two interactions. On the horizontal axis time is presented in seconds and on the vertical axis the likelihood of a listener response according to the model is presented. The gray horizontal line is the validated threshold obtained during the validation step.

frame. After sequencing and smoothing these prediction values one gets a prediction value curve. In Figure 10.1 two example prediction value curves are presented, plotted in black. These examples were taken from the first minute of two interactions from the MultiLis corpus and produced by the "Consensus 2" model from Chapter 7.

The timing predictions for listener responses were be extracted from these prediction value curves. This was done by detecting peaks in the prediction value curve and comparing these peaks to a fixed threshold. If this peak exceeded this fixed threshold, it was considered appropriate to give a listener response at the time of the peak. The fixed threshold was determined in the validation phase of the development of the model. This threshold can be decreased or increased to generate more or fewer responses respectively to express more attention or a different personality type. In Figure 10.1 the threshold that was found to give the highest $F_1$ score for the "Consensus 2" model is indicated by the horizontal line at 0.2122.

This is the way that other probabilistic models determined their timings as well [107, 79, 116]. In this Chapter a new approach to select the predictions from the prediction value curve will be presented. In Section 10.1 the limitations of the fixed threshold will be discussed. Our dynamic thresholding approach will be presented in

Section 10.2. The new dynamic thresholding approach will introduce a new problem which will be addressed in Section 10.3. The approaches will be evaluated objectively on the MultiLis corpus in Section 10.4 and using the individual perceptual evaluation method in Section 10.5.

## 10.1 Limitations of the Fixed Threshold

Selecting suitable moments in the prediction value curve for prediction using a fixed threshold has a problem. The number of listener responses predicted by the model using a certain threshold is inconsistent. The same model using the same threshold applied to two different speakers can result in a significant variation in listener response rate. This is problematic for the designer of an embodied conversational agent. The designer usually has a personality or role in mind for their agent. A lot of factors are important when managing the impression a user has about the personality of an agent. One of these factors is their listening behavior. The timing, number and form of the listener responses that are produced by an embodied conversational agent have been proven to influence the impression the user has about their personality [38]. So, the stability of the generated behavior is important.

That this stability is not provided by the current fixed threshold is illustrated by the two prediction value curves in Figure 10.1. Looking at the number of peaks that exceed the threshold shows that applying the optimal threshold according to the validation step has resulted in a big difference in response rate. For the first interaction nineteen listener responses are predicted in the first minute, while only one is predicted for the second interaction. The explanation for the lower prediction values in the bottom curve in Figure 10.1 lies in the fact that in this case the speaker hardly ever looked at the listener whilst gaze is one of the most important cues that the prediction model uses. Even though the speaker does not often look at the listener, opportunities to give a listener response are available, since the listeners in the corpus did respond during this minute. So, it might be in the interest of the virtual listener to give a listener response as well. Either to comfort and encourage the speaker to continue speaking and/or to build a better rapport with the speaker.

With a fixed threshold this is hard to do, since there is no reliable way of knowing how much one needs to lower the threshold to get the desired response rate. This is because selecting peaks based on a fixed validated threshold is subject to changing conditions. These changing conditions do not limit themselves to eye gaze behavior as in the previous example. Other aspects of a speaking style can also change. Examples include, but are not limited to speaking in a louder or softer voice, with higher or lower pitch, different speech rates or varying degrees in which intonation is used. These aspects can even change during an interaction with the same speaker, not necessarily because the speakers changes these characteristics of their speaking style, but they may also change their position relative to the microphone or video camera. All these things can influence the features used by the model, which in turn influence the prediction values returned by the model. For some conditions the general prediction value from a model will always be lower than for other conditions and the peaks in this curve may remain below the fixed validated threshold.

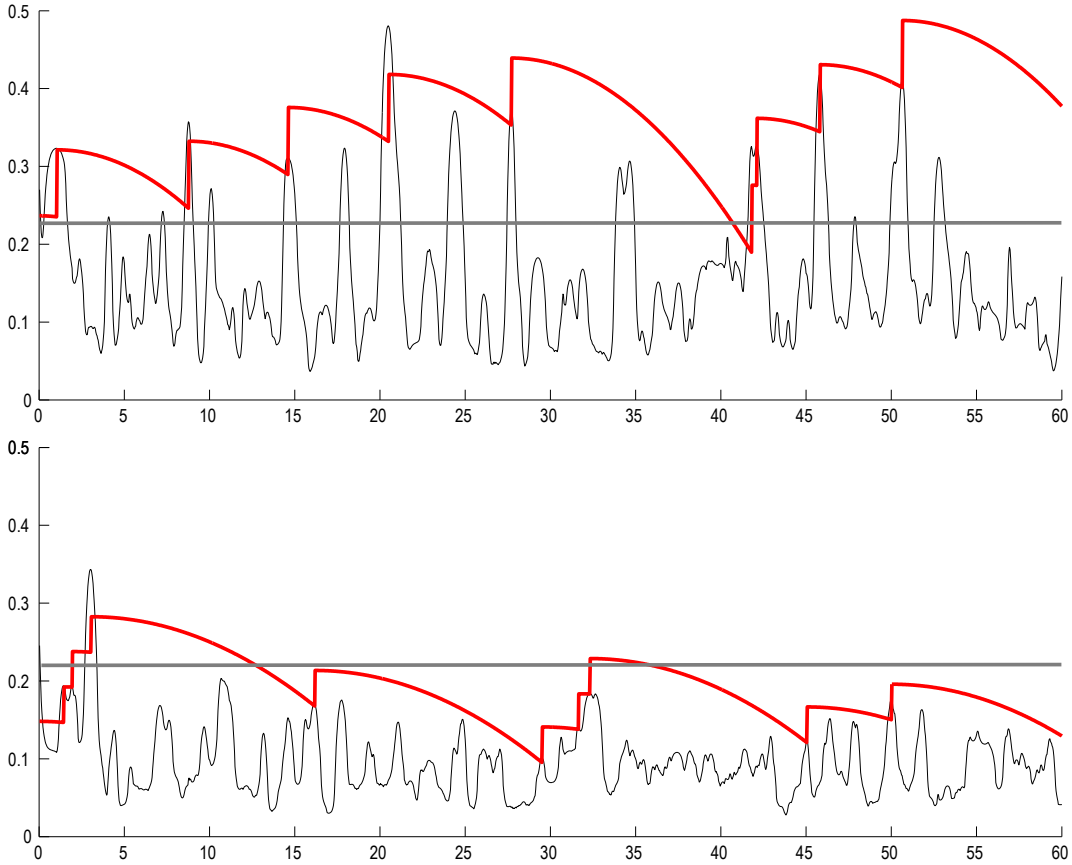


**Figure 10.2:** The proposed dynamic threshold is shown in red. The threshold starts initially high and decreases over time at an increasing rate until a peak exceeds the threshold. Then the threshold is increased by a fixed amount after which the threshold starts decreasing slowly again.

Essentially, these differing conditions make it very difficult for a designer of an embodied conversational agent to give the agent the personality and behavior they have in mind. For one user the agent may be responsive and attentive, while for another almost no listener responses are generated at all.

So, another way of extracting the appropriate timings of listener responses from the prediction value curves is needed. To give the designer of the nonverbal listening behavior of a virtual agent the tools to control the generated behavior, the solution needs to ensure the following characteristics:

- **Stable Response Rate** - The solution should be able to generate a similar response rate under changing conditions, such as different speakers, audio/video quality or feature extraction accuracy.
- **Evenly Distributed Responses** - Not only the overall rate should be stable, but the distribution needs to be even as well. Periods with many or few responses are perceived as unnatural behavior [123].
- **Adjustable Response Rate** - The response rate needs to be easily adjustable, so the designer of the agent can generate the desired behavior and even change this behavior during the interaction.

## 10.2 Dynamic Thresholding

The proposed approach changes the threshold dynamically during the interaction. So, instead of a fixed threshold determined at the development stage of the prediction model, a dynamic threshold is used that changes over time depending on the time since the last predicted listener response was proposed. At the start of the listening period the threshold is relatively high and it starts decreasing at an increasing rate until a peak in the curve exceeds the threshold. When a listener response is predicted the threshold jumps up and starts decreasing again at an initially slow rate. This will ensure both the stable response rate and the even distribution of the responses.

In Figure 10.2 the dynamic threshold for the two interactions is shown in red. By looking at the peaks that exceed the red threshold, the figure shows that for both example interactions the number of predicted response opportunities is nine. So, the resulting response rate is much more stable than the response rate the fixed threshold would have given, meaning we have met our first required characteristic. The responses are also more evenly distributed. There are no long periods without responses, the longest gap being 14 seconds. There are still issues with two or three consecutive predicted response opportunities, but an extra rule stating no two listener responses to be generated within a certain amount of time would solve this. For the final required characteristic a closer look at the formula that created this dynamic threshold is required.

$$T_t = T_{t_{last}} + j - (\frac{t - t_{last}}{g.r})^2.\frac{j}{d} \tag{10.1}$$

The formula that created the dynamic thresholds in red in Figure 10.2 is presented in Equation 10.1. It calculates the dynamic threshold ($T$) at time $t$. Time $t$ is measured in frames. Gap parameter $g$ is the mean time in seconds between two predicted listener responses. Parameter $r$ is the sampling rate of the system, needed to convert timing in seconds into timing in frames. Parameter $t_{last}$ is the time of the listener response that was last generated and $T_{t_{last}}$ is the dynamic threshold at that time. Parameter $j$ is the jump parameter, which represents the amount that the dynamic threshold increases after predicting a listener response. The final parameter $d$ is the drop off parameter, which controls the amount the dynamic thresholds decreases over time.

Gap parameter $g$ is the parameter that will give the designer of the agent control over the response rate of the agent by defining the mean gap between two predicted listener responses. However, before this gap parameter will give the expected behavior, the jump parameter $j$ and drop off parameter $d$ need to be determined. For the jump parameter $j$ we recommend using the standard deviation of the prediction value curve as value. This will make the jumps after predicting a listener response appropriate to the variation found in the prediction value curve and thus make the dynamic threshold even more adaptable to differing conditions. That leaves the drop off parameter and this needs to be calibrated on a development set of example interactions. The procedure for this is to try different combinations of parameter $d$ and $g$ and minimize the absolute difference in expected number of listener responses and predicted number of listener responses for the values of $g$ you expect the virtual listener to use.

## 10.3 Variable Head Nods

The application of the dynamic threshold comes with a cost though. The dynamic threshold forces a certain response rate on an interaction. To uphold the requested response rate, some of the peaks that get selected by the declining threshold have a low(er) prediction value and are thus less likely to be good places to give a listener response. There are several ways one could deal with this. For instance, one could introduce another threshold below which no listener responses should be generated. Here another possibility involving a change in the type of response that is generated is explored.

Since the model is less confident about the correct timing of these listener responses, it will possibly be better to make them less noticable. This listener response will most likely be interpreted more as a signal of attention instead of an acknowledgment of given information and correct timing of such signals is less essential. So, even if the smaller listener response is not timed correctly, the interaction is unlikely to break down from it.

To achieve this, the height of the peak in the prediction value curve is used as measure for the amplitude of the generated head nod. So, instead of generating the same head nod at each predicted response opportunity, a head nod with a larger amplitude is generated when the peak in the prediction value curve is high and one with a smaller amplitude when the peak in the prediction value curve is low.

## 10.4 Objective Evaluation

In Section 10.2 the dynamic threshold formula was presented and its merits were highlighted on two segments of one minute. In this section the performance of the dynamic threshold over a larger sample size will be evaluated. The procedure of this evaluation will be explained in Section 10.4.1.

To support the dynamic thresholding formula and the choice to generate variable head nods based on the height of the peaks, the evaluation will aim to give answers to the following questions: Does the dynamic threshold succeed in stabilizing the response rate? Does it avoid periods with few or many responses? Are the lower peaks that get selected by the dynamic threshold indeed the worst times to give a listener response?

### 10.4.1 Procedure

For the evaluation the “Consensus 2” model from Chapter 7 was used. This model was applied to the ten interactions from the test set from that experiment. Thus, these interactions were not used in the development of the model. The listener response prediction model was applied to these ten interactions to obtain the prediction value curve for each interaction. Then eleven fixed and eleven dynamic thresholds were applied to these prediction value curves. The fixed thresholds were varied between 0.15 and 0.35.

The eleven dynamic thresholds were selected such that the resulting overall response rate was similar to the mean response rate of each of the fixed thresholds.

| | Response Rate for Fixed Threshold (responses/minute) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Interaction | | | | | | | | | |
| **Threshold** | **Overall** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **0.15** | 21.6 | 19.0 | 23.8 | 26.1 | 21.9 | 12.9 | 21.8 | 24.9 | 24.2 | 22.7 | 21.5 |
| **0.17** | 16.8 | 15.7 | 20.2 | 21.6 | 17.2 | 6.4 | 16.7 | 22.3 | 19.2 | 16.4 | 16.1 |
| **0.19** | 13.2 | 13.7 | 16.2 | 18.6 | 12.1 | 3.0 | 12.0 | 18.4 | 15.5 | 13.4 | 13.1 |
| **0.21** | 11.5 | 12.8 | 14.8 | 16.7 | 11.1 | 2.0 | 8.8 | 15.3 | 13.0 | 12.5 | 11.5 |
| **0.23** | 10.1 | 11.3 | 13.3 | 15.7 | 10.0 | 1.5 | 7.2 | 14.4 | 11.3 | 10.8 | 10.9 |
| **0.25** | 8.7 | 9.3 | 11.5 | 14.1 | 8.7 | 1.0 | 5.5 | 13.6 | 9.8 | 9.6 | 8.4 |
| **0.27** | 7.1 | 7.0 | 10.4 | 13.4 | 8.0 | 1.0 | 3.5 | 11.8 | 8.2 | 7.2 | 6.6 |
| **0.29** | 5.5 | 4.0 | 8.6 | 12.4 | 6.7 | 1.0 | 2.9 | 10.9 | 6.3 | 5.3 | 4.1 |
| **0.31** | 4.4 | 3.1 | 6.1 | 10.1 | 5.4 | 0.7 | 2.1 | 10.1 | 5.3 | 3.8 | 3.2 |
| **0.33** | 3.1 | 2.2 | 4.3 | 5.9 | 4.1 | 0.7 | 1.9 | 8.7 | 2.8 | 2.8 | 2.7 |
| **0.35** | 2.4 | 1.8 | 2.5 | 4.9 | 3.1 | 0.5 | 1.5 | 6.6 | 2.3 | 1.3 | 2.5 |

**Table 10.1:** The table illustrates the effect of different fixed thresholds on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, although increasing the threshold decreases the overall response rate at a predictable pace, the effects on individual interactions varies wildly.

Thus, the gap parameter $g$ from the dynamic threshold formula was set such that the response rate of that threshold was (almost) the same as the mean result rate of the corresponding fixed threshold. For both the fixed and the dynamic threshold, listener responses were not allowed to appear within one second of the previous listener response. These predicted listener responses were discarded.

For the dynamic threshold other parameters beside the gap parameter $g$ needed to be set as well. For this evaluation the dynamic threshold were initialized with the mean of the prediction value curve as initial $T_{t_{last}}$ and the standard deviation of the prediction value curve as jump parameter $j$. To select the value for the drop off parameter $d$, several combinations of gap parameter $g$ and drop off paramater $d$ were tried on the ten interactions. Using this data the value for drop off parameter $d$ was set such that the difference between the expected number of listener responses based on the value for gap parameter $g$ and the resulting number of predicted listener responses was minimized. This was true for value 1.4.

### 10.4.2 Results

The first question that will be answered is, does the dynamic threshold succeed in stabilizing the response rate? For this a comparison was made between the response rates that were the result of applying the eleven fixed and dynamic thresholds on each interaction. These response rates in responses per minute are presented in Tables 10.1 (fixed thresholds) and Table 10.2 (dynamic thresholds). In the first column the height of the fixed threshold and the gap parameter $g$ that were varied are presented. A comparison of the second columns of Table 10.1 and Table 10.2 shows that a similar overall response rate is predicted by the model by both thresholds.

In the remaining columns the response rates for each interactions are presented. The cells are shaded gray for easier interpretation, with higher response rate being

| | **Response Rate for Dynamic Threshold** (responses/minute) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Interaction** | | | | | | | | | |
| **Gap** $g$ | **Overall** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **2.8** | 18.1 | 17.7 | 19.1 | 18.6 | 18.0 | 17.8 | 18.4 | 19.2 | 17.0 | 18.5 | 18.1 |
| **3.6** | 14.8 | 14.1 | 14.8 | 15.4 | 14.7 | 14.4 | 14.6 | 15.7 | 14.5 | 15.1 | 15.2 |
| **4.5** | 12.1 | 11.9 | 12.2 | 12.4 | 12.1 | 11.9 | 11.9 | 13.1 | 12.0 | 12.1 | 12.2 |
| **5.2** | 10.6 | 10.4 | 10.4 | 10.8 | 10.5 | 10.2 | 11.0 | 11.4 | 10.5 | 10.8 | 10.6 |
| **5.9** | 9.5 | 9.8 | 9.4 | 10.1 | 9.3 | 8.9 | 9.3 | 10.1 | 9.2 | 9.8 | 9.9 |
| **6.9** | 8.4 | 8.8 | 8.3 | 8.5 | 8.0 | 7.9 | 8.3 | 9.2 | 8.2 | 8.7 | 8.6 |
| **8.4** | 7.0 | 6.8 | 7.6 | 7.5 | 6.9 | 6.7 | 6.7 | 7.9 | 6.7 | 7.4 | 7.2 |
| **10.9** | 5.7 | 5.6 | 5.8 | 6.2 | 5.9 | 5.2 | 5.5 | 6.6 | 5.3 | 5.7 | 5.7 |
| **13.7** | 4.7 | 4.3 | 4.3 | 5.2 | 4.9 | 4.7 | 4.5 | 5.2 | 4.5 | 4.9 | 4.7 |
| **19.2** | 3.6 | 3.3 | 4.0 | 3.6 | 3.6 | 3.5 | 3.4 | 3.9 | 3.3 | 3.8 | 3.6 |
| **25.2** | 2.9 | 2.6 | 3.2 | 3.1 | 3.1 | 3.0 | 2.8 | 3.5 | 2.7 | 3.2 | 2.9 |

**Table 10.2:** The table illustrates the effect of different gap parameters in the dynamic threshold on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, for each interaction a similar response rate is generated.

darker. These rates show that for the fixed threshold the response rate of each interaction can vary wildly. Especially the predicted response rate for interaction 5 is a lot lower than for the other interactions, while the predicted response rate for interaction 7 is generally higher. For the dynamic threshold the response rates for each individual interaction are a lot more stable. Each interaction has more or less the same response rate. So indeed, the dynamic threshold has succeeded in stabilizing the response rate.

The next question that will be addressed is, does the dynamic threshold avoid periods with many or few responses? First, we analyzed whether the dynamic threshold avoids periods with many responses. As a measure for this the number of listener responses that were discarded by the extra rule that no predicted responses should be within a second of each other was used. The number of discarded predictions is presented in Table 10.3 and the left graph in Figure 10.3. They show that this number is about twice as high in the case of the fixed threshold compared to the dynamic threshold. So, the dynamic threshold helps to avoid periods with many responses.

To see whether the dynamic threshold helps to avoid periods with few predicted responses, an analysis of the gaps between two consecutive predicted responses was performed. As a measure the mean, standard deviation and maximum of these gaps was used. The results of this analysis are presented in Table 10.3. Comparing columns $\mu$ and max of both thresholds shows that the difference between the mean and maximum gap is a lot more stable for the dynamic threshold (see also right graph in Figure 10.3). This is also reflected in the lower standard deviation for the dynamic threshold (see middle graph in Figure 10.3). So, it can be concluded that the dynamic threshold also resulted in a more even distribution of the predicted responses.

Previously in Section 10.3 it was argued that the lower peaks in the prediction value curves were less likely to be good places for listener responses. Although this is intuitively correct no proof was provided to support this claim. To demonstrate that this is indeed the case, an analysis was performed of the peaks that were selected by the eleven different dynamic thresholds. These peaks were divided into two groups

| Fixed Threshold | | | | | Dynamic Threshold | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Discarded | Gap (s) | | | | Discarded | Gap (s) | | |
| Threshold | Predictions | $\mu$ | $\sigma$ | max | Gap $g$ | Predictions | $\mu$ | $\sigma$ | max |
| **0.15** | 384 | 2.75 | 1.85 | 24.17 | **2.8** | 129 | 3.31 | 1.41 | 7.33 |
| **0.17** | 202 | 3.56 | 2.93 | 47.10 | **3.6** | 75 | 4.06 | 1.89 | 9.37 |
| **0.19** | 114 | 4.51 | 5.12 | 85.63 | **4.5** | 52 | 4.97 | 2.53 | 10.93 |
| **0.21** | 79 | 5.21 | 6.06 | 85.63 | **5.2** | 43 | 5.65 | 2.98 | 13.67 |
| **0.23** | 62 | 5.91 | 6.68 | 85.63 | **5.9** | 40 | 6.35 | 3.42 | 15.17 |
| **0.25** | 55 | 6.94 | 8.25 | 97.60 | **6.9** | 27 | 7.17 | 4.09 | 17.13 |
| **0.27** | 41 | 8.51 | 9.60 | 97.60 | **8.4** | 18 | 8.61 | 4.95 | 20.97 |
| **0.29** | 23 | 11.01 | 12.48 | 97.60 | **10.9** | 16 | 10.74 | 6.79 | 27.03 |
| **0.31** | 16 | 13.73 | 15.56 | 97.60 | **13.7** | 11 | 13.13 | 8.30 | 33.47 |
| **0.33** | 11 | 19.15 | 19.77 | 101.93 | **19.2** | 6 | 17.53 | 12.13 | 43.63 |
| **0.35** | 9 | 22.80 | 22.61 | 101.93 | **25.2** | 5 | 21.49 | 16.11 | 55.87 |

**Table 10.3:** Presentation of the results of the analysis of the gaps between predicted responses using a fixed threshold and our dynamic threshold.

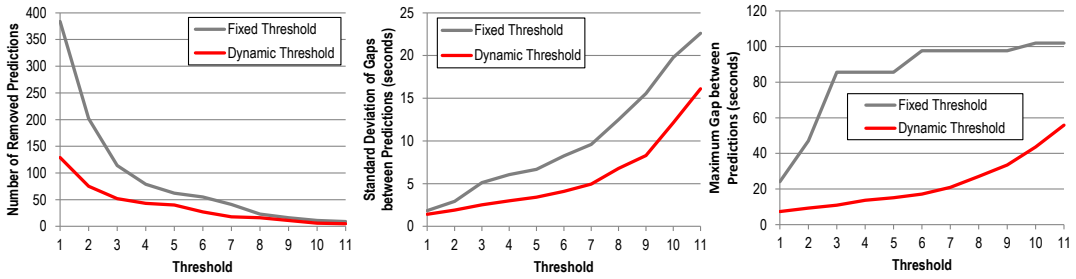


**Figure 10.3:** Three graphs to illustrate that for all thresholds the dynamic thresholding needs fewer predictions to be removed (left graph), has a lower standard deviation for the gaps between two consecutive predictions (middle graph) and has a smaller maximum gap between consecutive predictions (right graph).

and we checked whether a listener response could be found from one of the recorded listeners in the MultiLis corpus. The two groups were: peaks *above* the fixed threshold that results in the best $F_1$-score in the experiments to evaluate the model, namely 0.2122 (see Chapter 7), and peaks *below* this fixed threshold. For the peaks *above* the threshold a listener response could be found for 53% of the peaks (1666 out of 3149), while for peaks *below* the threshold this was only true for 15% of the peaks (176 out of 1156). So, indeed the lower peaks are less likely to be good places for a listener response and measures to counteract this effect of the dynamic thresholding need to be taken.

## 10.5 Subjective Evaluation

So far it was objectively established that dynamic thresholding succeeds in stabilizing the response rate. But this does not necessarily mean that the generated behavior is also acceptable for humans. The response rate was stabilized by selecting suboptimal peaks in the prediction value curve and generating listener responses at those peaks. The previous evaluation showed that at those suboptimal peaks the listeners in the MultiLis corpus responded fewer times at those moments than at the higher peaks.

The question is whether the listener responses generated at these lower peaks influence the ratings of the listening behavior and whether adjusting the amplitude of the head nods has a positive effect on this.

To answer these questions a perception experiment was set up in which participants were asked to select their preferred listener out of two generated listeners. The procedure of the experiments will be explained in more detail in Section 10.5.1, the stimuli in Section 10.5.2. The measures that were used in our evaluations will be explained in Section 10.5.3. The results will be presented and discussed in Section 10.5.4.

### 10.5.1 Procedure

For the evaluation 17 participants (3 female, mean age 27) were recruited among colleagues and students. The individual perception evaluation method was used. The experiment was presented through a webpage. After a short introduction on what listening behavior is and what the function of the behavior is, the participants watched four pairs of two interactions between a recorded speaker from the MultiLis corpus and a generated virtual listener. The participants were shown the same interaction twice, but each time with a different generated listener. The difference between the two generated listeners was either the threshold used to determine the number and timing of the listener responses (twice) or the variation in amplitude of the head nods based on confidence of the prediction model (twice). To eliminate learning effects on our results the order of the stimuli were systematically varied between participants.

While watching each interaction, the participants were instructed to pay attention to the timing of the listener responses and judge each listener response on whether or not they thought the listener response was appropriately timed. When a listener response was inappropriate according to their judgment they were instructed to press the spacebar (a yuck response). The participants had the option to replay the video, which would result in a loss of all collected judgments for that video so far.

After each pair of stimuli of the same interaction, they were asked which of the two generated virtual listeners they preferred.

### 10.5.2 Stimuli

The stimuli that were presented to the participants were side-by-side interactions between a speaker from the MultiLis corpus and a generated virtual listener (see Figure 10.4). The speakers were the speakers from the same ten interactions used in the objective evaluation. In response to these speakers we generated virtual listeners using the BML realizer Elckerlyc [142]. There were four versions of these virtual listeners generated for each interaction.

Each virtual listener was generated using the *fixed* or *dynamic* threshold and with the *same* head nod each time or with *variable* head nods. For the fixed threshold the threshold that was validated to achieve the highest $F_1$ score in the original paper (0.2122) was used. For the values for the parameters of the dynamic threshold the same procedure as explained in Section 10.4.1 was used. Gap parameter $g$ was set to 7, which corresponds to the mean response rate found in the MultiLis corpus.

**Figure 10.4:** Screenshot of one of the stimuli used in the perception study.

At each selected peak a head nod was generated. The behavior was realized with the BML element <head>. Specifically, a head nod was generated that started downward, lasted 0.4 seconds, reached its lowest point after 0.13 and stayed there for 0.02 seconds. For the listeners with the same head nod each time, the amplitude of the behavior was set to 0.15. For the listeners with variable head nods, the amplitude corresponded to half of the prediction value of the peak in the prediction value curve. This resulted in a minimum amplitude of 0.04 and a maximum of 0.29.

### 10.5.3 Measures

The preference of the participants was measured by an objective and a subjective method.

For the objective method the number of yuck responses was recorded during the interaction. Because of the variation in response rate between the fixed and dynamic threshold the number of the yuck responses was divided by the number of judged head nods. The listener with the lowest number of yuck responses per judged head nod was counted as the preferred listener.

The subjective preference of the participant was recorded by a slider with the first listener on the left end and the second listener on the right end. The slider was initialized in the middle (no preference). Furthermore, the participants were asked what they thought to be the difference between the two generated listeners through an open question to gain insight into which characteristics they based their decision on.

### 10.5.4 Results

In the experiment two comparisons were made by the participants: a comparison between *fixed* and *dynamic* threshold and a comparison between the *same* and *variable* head nods.

The results for the comparison between the fixed and dynamic threshold are presented in Figure 10.5. The height of the bars in the figure shows the number of times the listener generated using a fixed or dynamic threshold was preferred by the participants based on the subjective preference rating (left) or the objective measure

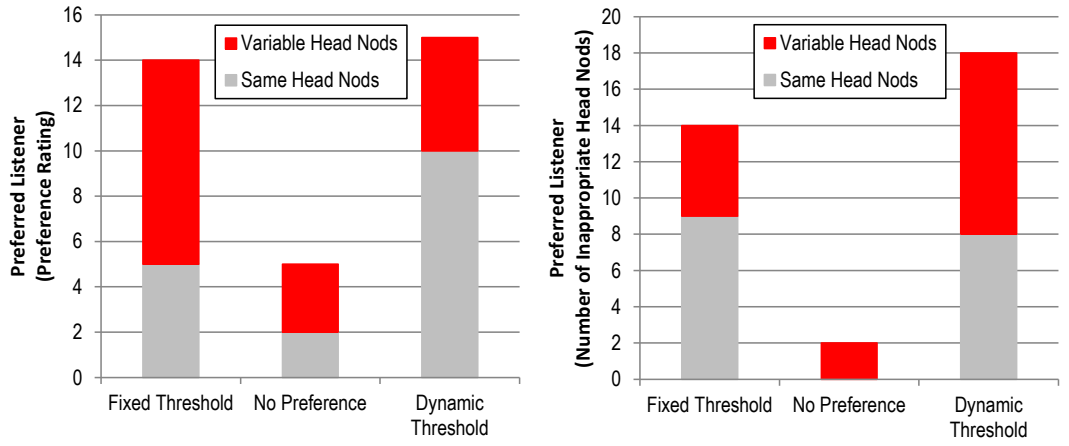


**Figure 10.5:** The height of the bars shows the number of times the listener generated using a fixed or dynamic threshold was preferred by the participants based on preference rating (left) or number of inappropriate head nods (right).

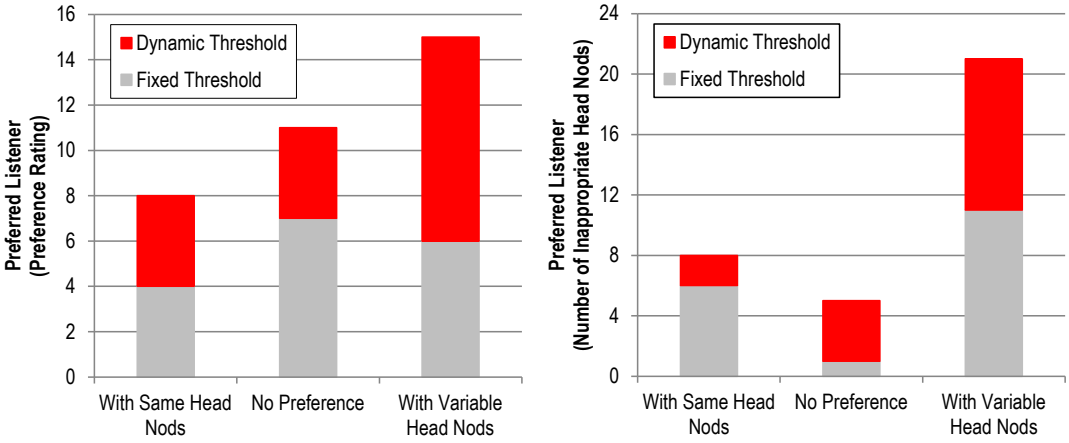


**Figure 10.6:** The height of the bars shows the number of times the listener with the same head nods or variable head nods was preferred by the participants based on preference rating (left) or number of inappropriate head nods (right).

of number of inappropriate head nods per judged head nod (right). The results are polarizing. For the subjective measure based on the preference ratings from the participants (left graph) the listener using dynamic thresholding was selected once more (15 versus 14 for fixed thresholding) and 5 times no preference was indicated. Based on the objective measure of inappropriate judgments per judged head nod the difference is a little bit bigger. The dynamic thresholding had 18 times the least number of inappropriate judgments per judged head nod and fixed thresholding 14 times. Two times this rate was the same.

So, at first sight there does not appear to be a clear favorite. Looking at the answers to the open question about the motivation for their choice reveals that oftentimes the decision was based on the number of listener responses given. If this was the motiviation, most often the listener with the most responses was selected as

the preferred listener. In the stimuli there was an imbalance between the two conditions in this regard. For nine out of ten interactions the listener generated with the fixed thresholds had more responses than the version with the dynamic threshold. In total 495 responses were generated in the fixed threshold condition and 366 in the dynamic threshold condition. Therefore, there was an unexpected bias in our stimuli towards the fixed threshold. Despite this bias, the dynamic threshold still comes out slightly ahead or at least equal to the fixed threshold.

The results for the comparison between the same and variable head nods are presented in Figure 10.6. The height of the bars in the figure shows the number of times the listener generated using the same or variable head nods was preferred by the participants based on the subjective preference rating (left) or the objective measure of number of yuck responses per judged head nod (right). Based on the preference rating (left graph) the variable head nods were more often preferred than the same head nods (15 times versus 8 times, with 11 times no preference). Based on the objective measure this difference was even bigger (21 times for variable head nods, 8 times for the same head nods and 5 times no preference).

The higher number of no preference ratings shows that the participants had a harder time choosing between these two versions of the virtual listener. Based on the open question only one participant seemed to have consciously noticed the varying amplitude of the generated head nods and has provided this as a reason for his/her preference. This is not a surprise, since the participants were not instructed to pay attention to the form of the behavior. During the interaction they were mainly focussed on the timing of the listener responses and hitting the keyboard any time they saw a head nod that was inappropriately timed. However, the fact that the amplitude of the head nod factored into the decision to judge a head nod as inappropriate or not is likely. This is shown by the fact that the listeners in variable head nods condition had fewer yuck responses. The number of times the participants hit the keyboard during an interaction subsequently carried over into their decision on their preference.

## 10.6 Conclusion

In this Chapter we have presented a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models in predicting generic listener responses for embodied conversational agents. Based on the time since the last generated listener response the dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to create the behavior that is desired for the targetted role, personality or situation.

Through objective and subjective evaluation it was shown that the generated behavior was more stable under changing conditions and was perceived to be at least equally humanlike to the behavior of the traditional fixed threshold. Furthermore, it was shown that the “strength” of the listener response can be varied using the height of the selected peak in the prediction value curve and that this is preferred over gen-

erating the same head nod all the time.

# Part IV

# Concluding Thoughts

# 11

# Reflection and Future Work

One of the questions to address research on spoken dialogue systems and embodied conversational agents is to have the system produce appropriate behavior when the user of the system is speaking. In human-human conversations, listeners produce listener responses to indicate that they are hearing, understanding and assessing what the speaker is saying. This behavior is a key part of the grounding process which is one of the structures in human-human communication.

In the computational literature there is a line of research that tries to build artificial listeners that generate such responses in a reactive manner: without intricate deliberation and only based on superficial cues from the speaker's voice, sometimes combined with visual information. The computational models that are built in this line of research are based on the analysis of and learning on recordings of human-human recordings.

In this thesis we have expanded on this line of research. We have analyzed and built reactive models for a subset of listener responses, namely generic listener responses. The presented work was executed under the assumption is that there are some places where a generic listener response is more appropriate (perhaps even mandatory) and places where it would be perceived strange if the system would produce a generic listener response and that these places can be extracted from these human-human recordings. These different type of places, which we have called responses opportunities, were analyzed in human-human recordings and utilized for modeling in several ways - introducing new methods of data collection to get more insight in the characteristics of the response opportunities and new strategies to use the data to build reactive models of listener responses.

The contributions of this thesis can be summarized by the following:

- **The Concept of Response Opportunities** - Central to the work presented in this thesis was the concept of *response opportunities*. A response opportunity was defined as a window in time where a listener response is appropriate. The response opportunities were identified by collecting multiple perspectives on appropriate and inappropriate listening behavior. With each response opportunity comes the number of listeners that responded to that response opportunity.

Conversational analysis showed differences between important response opportunities and less important response opportunities. The concept of response opportunities and the associated importance of the response opportunity proved to be a useful concept to improve the learning and evaluating of listener response prediction models.

- **Methods for Collecting Multiple Perspectives on Listening Behavior** - To identify as many response opportunities in an interaction as possible we presented three methods to collect multiple perspectives on listening behavior. Through a unique parallel recording setup the MultiLis corpus captured three listeners in simultaneous interaction with the same speaker. These three listeners gave three perspectives on appropriate listening behavior. The second method, parasocial sampling, proved to be a suitable substitute to collect additional perspectives on an existing corpus. Furthermore, individual perceptual evaluation was introduced to collect perspectives on inappropriate moments to give a listener response. These perspectives can be combined in a consensus perspective, which gives an overview of the response opportunities and inappropriate moments for listener responses for an interaction.

- **Conversational Analysis of the Characteristics of Response Opportunities** - Conversational analysis of the response opportunities was performed looking into the relation between actions of the speaker and response opportunities. The analysis was performed on the eye gaze of the speaker, pitch and energy of the speech signal and the type of statements that the listener responded to. This analysis showed that the cues known to be associated with listener responses, such as the speaker looking at the listener and the end of an utterance, are more often present at the response opportunities where the majority of the listeners responded than the response opportunities where only one listener responded. Furthermore, the conversational analysis showed that summarizing and repeating statements more often elicited a response from all listeners and that listeners were more likely to deviate from their typical response at such times.

- **Methods for Learning Listener Response Prediction Models using Differences and Similarities between Interlocutors** - The thesis presented three novel methods to improve the learning of listener response prediction models. Since each of the presented methods focussed on improving a different aspect of the learning process they could be combined. The first method presented a way to collect more accurate positive ground truth samples. This was done by using only the response opportunities where the majority of the persepectives responded as positive ground truth samples. The second method presented a way to collect more accurate negative ground truth samples. This method revolved around using the subjective ratings of observers as negative ground truth samples in an iterative learning setup. The final method presented a way to split the learning data to learn a speaker-adaptive listener response prediction model. This method learned a model for each interaction individually, selected the best performing models for the model collection and matched new speakers with the speakers from the learning data to select the appropriate model from the model

collection.

Furthermore, the thesis presented methods to improve the evaluation of listener response prediction models with a new performance objective measure $F_{consensus}$ and the subjective individual perceptual evaluation method. As a final contribution the dynamic thresholding technique was presented as a method to generate personalizable and predictable listening behavior for embodied conversational agents using listener response prediction models.

## 11.1 Limitations and Future Work

All analyses in this thesis were performed on the MultiLis corpus. Due to the parallel recording setup the interaction in the MultiLis corpus was restricted. No turn-taking was allowed and the flow of information was one-sided, from the speaker to the listener. This brings certain characteristics to the interactions which may not translate to more interactive conversational settings. The question is whether the results presented in this thesis, whether it be the results of the conversational analysis or the performances of the listener response prediction models, transfer over to conversations where turn-taking behavior becomes involved. Are the cues for response opportunities the same in less structured and more interactive interactions than the interaction with an information giving task used for the MultiLis corpus? And consequently, do the listener response prediction models also work for such interactions? The assumption in this thesis is that they do, but without comparative studies this is open for debate.

The listener response prediction models developed in this thesis only model the timing of *generic* listener responses. These generic listener responses have the function to signal attendence and a general notion of understanding to let the speaker know he/she can continue. Typical examples of such generic listener responses include nods and minimal verbal utterance such as “mm-hmm” or “yeah”. This is only a subset of the behavior that listeners display and additional studies are required to built models that are capable of capture the complete listening behavior of humans.

Furthermore, the listener response prediction models presented in this thesis work on features that only represent superficial behavior. The presented models perform at an adequate level based on these features (for instance only 8% of the generated responses of the subjective evaluation presented in Section 10.5 were found to be inappropriate), but more advanced features representing the meaning and level of understanding of the contributions of the speaker are needed to advance the state-of-the art of such models. Work towards such features exists [39, 140], but so far such features have not been used for listener response prediction models.

Also other aspects of the listener response prediction models can be improved upon. Development of such models should move away from building one general model representing the consensus of all speakers and listeners in a corpus. Each human is unique in their behavior and adaptation is something that is one of the fundamental processes in conversation. The speaker-adaptive listener response prediction model presented in Chapter 9 is a promising step into this direction. The working of the current speaker-adapative model has room for improvement. The approach

would probably benefit from more sophisticated speaker descriptors than currently used. The conversational analysis results from Part II showed that listener responses were usually placed near the end of an utterance. The current speaker descriptors summarize the behavior of the speakers during the whole interaction, but more targetted descriptors for the behavior near the end of an utterance are possibly more informative for finding the model that is trained on the most similar speaker.

Integration of the listener response prediction models in an interactive virtual agent was originally one of the goals. The dynamic threshold and variable head nods have been proposed as tools for integration. However, many other aspects still need further investigation. For instance, currently the model only generates head nods as listener responses, but vocal listener responses are also common.

Ultimately, listening behavior is only a small portion of the behavior of humans during conversations. The models proposed in this thesis need to be integrated with models for speaking behavior and models for turn-taking behavior which handles the transition between these two conversational roles. And these are only some of the challenges that lie ahead before the interaction with an embodied conversational agent feels like an interaction with a human. By using and refining the methods for data collection that were introduced and used in this thesis and applying them to other issues related to human-agent communication, these challenges will hopefully become less of a challenge.

# Publications

## Included in this Thesis


Iwan de Kok, Derya Ozkan, Dirk Heylen and Louis-Philippe Morency. Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 3-10, 2010.

Iwan de Kok and Dirk Heylen. The MultiLis Corpus - Dealing with Individual Differences in Nonverbal Listening Behavior, in *Proceedings of the Third COST 2102 International Training School*, pages 362-375, 2011.

Iwan de Kok and Dirk Heylen. Observations on Listener Responses from Multiple Perspectives, in *Proceedings of the 3rd Nordic Symposium on Multimodal Communication*, pages 48-55, 2011.

Iwan de Kok and Dirk Heylen. Appropriate and Inappropriate Timing of Listener Responses from Multiple Perspectives, in *Intelligent Virtual Agents*, pages 248-254, 2011.

Iwan de Kok and Dirk Heylen. Analyzing Nonverbal Listener Responses using Parallel Recordings of Multiple Listeners. *Cognitive Processing*, 13(2):499-506, 2012.

Iwan de Kok and Dirk Heylen. A survey on evaluation metrics for backchannel prediction models, in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, pages 15-18, 2012.

Iwan de Kok and Dirk Heylen. Integrating Backchannel Prediction Models into Embodied Conversational Agents. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents (IVA 2012)*, pages 268-274, 2012.

Iwan de Kok and Dirk Heylen. Controlling the Listener Response Rate of Virtual Agents. Accepted for publication at *the 12th International Conference on Intelligent Virtual Agents (IVA 2013)*, 2013.

Iwan de Kok, Dirk Heylen and Louis-Philippe Morency. Speaker-Adaptive Multimodal Prediction Model for Listener Responses. Submitted to *the 15th ACM International Conference on Multimodal Interaction (ICMI 2013)*, under review.

Iwan de Kok and Dirk Heylen. Generating Generic Listener Responses for Virtual Agents. *under review*.

Iwan de Kok, Ronald Poppe and Dirk Heylen. Iterative Perceptual Learning for Social Behavior Synthesis, *under review*.

## Additional Publications

Louis-Philippe Morency, Iwan de Kok and Jonathan Gratch. Context-based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 181-188, 2008.

Louis-Philippe Morency, Iwan de Kok and Jonathan Gratch. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *Intelligent Virtual Agents*, pages 176-190, 2008.

Iwan de Kok and Dirk Heylen. Multimodal End-of-Turn Prediction in Multi-Party Meetings. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 91-98, 2009.

Louis-Philippe Morency, Iwan de Kok and Jonathan Gratch. A Probabilistic Multimodal Approach for Predicting Listener Backchannels, *Journal of Autonomous Agents and Multi-Agent Systems*, 20(1):70-85, 2010.

Iwan de Kok and Dirk Heylen. Differences in Listener Responses between Procedural and Narrative Tasks, in *Proceedings of the 2nd International Workshop on Social Signal Processing*, pages 5-10, 2010.

Rieks op den Akker, Mariet Theune, Khiet Truong and Iwan de Kok. The Organisation of Floor in Meetings and the Relation with Speaker Addressee Patterns, in *Proceedings of the 2nd International Workshop on Social Signal Processing*, pages 35-40, 2010.

Dennis Reidsma, Iwan de Kok, Daniel Neiberg, Satish Pammi, Bart van Straalen, Khiet Truong and Herwin van Welbergen. Continuous Interaction with a Virtual Human. Journal on Multimodal User Interfaces, 4(2):97-118, 2011.

Khiet Truong, Ronald Poppe, Iwan de Kok and Dirk Heylen. A Multimodal Analysis of Vocal and Visual Backchannels in Spontaneous Dialogs. In *Proceedings of Interspeech*, pages 2973-2976, 2011.

Iwan de Kok and Dirk Heylen. When Do We Smile? Analysis and Modeling of the Nonverbal Context of Listener Smiles in Conversation. In *Proceedings of the 4th 4th International Conference on Affective Computing and Intelligent Interaction*, pages 477-486, 2011.

# Bibliography


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

[2] Jens Allwood and Loredana Cerrato. A study of gestural feedback expressions. In *Proc. of the First Nordic Symposium on Multi-modal Communication*, pages 7–20, 2003.

[3] Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, 9(1):1–26, 1992.

[4] Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366, 1991.

[5] Michael Argyle and Mark Cook. *Gaze and mutual gaze*. Cambridge University Press, London, United Kingdom, 1976.

[6] Michael Argyle, Roger Ingham, Florisse Alkema, and Margaret McCallin. The different functions of gaze. *Semiotica*, 7(1):19–32, 1973.

[7] Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[8] Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580, 2002.

[9] Janet Beavin Bavelas and Jennifer Gerwing. The Listener as Addressee in Face-to-Face Dialogue. *International Journal of Listening*, 25(3):178–198, 2011.

[10] Stefan Benus, Agustín Gravano, and Julia Hirschberg. The prosody of backchannels in american english. In *In Proceedings of ICPhS*, pages 1065–1068, 2007.

[11] Roxane Bertrand, Morgane Ader, Philippe Blache, Gaëlle Ferré, Robert Espesser, and Stéphane Rauzy. Représentation, édition et exploitation de données multimodales : le cas des backchannels du corpus CID. *Cahiers de l'Institut de Linguistique de Louvain*, 33(2):1–18, 2009.

[12] Roxane Bertrand, Philippe Blache, and Gaëlle Ferré. Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Languages*, 49(3), 2008.

[13] Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. Backchannels revisited from a multimodal perspective. In *Proceedings of Auditory-visual Speech Processing*, pages 1–5, 2007.

[14] Camiel J Beukeboom. When words feel right : How affective expressions of listeners change a speaker's language use. *European Journal of Social Psychology*, 756:747–756, 2009.

[15] Elisabetta Bevacqua, Dirk Heylen, Catherine Pelachaud, and Marion Tellier. Facial feedback signals for ecas. In *Proceedings of AISB*, 2007.

[16] Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agents*, pages 262–269, 2008.

[17] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. Multimodal Backchannels for Embodied Conversational Agents. In *Intelligent Virtual Agents*, pages 194–200, 2010.

[18] Susan E. Brennan and Eric A. Hulteen. Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems*, 8(2-3):143–151, June 1995.

[19] Susan E. Brennan and Michael F. Schrober. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44(2):274–296, 2001.

[20] Hennie Brugman and Albert Russel. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2065–2068, 2004.

[21] Lawrence J. Brunner. Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5):728–734, 1979.

[22] Hendrik Buschmeier and Stefan Kopp. Towards conversational agents that attend to and adapt to communicative user feedback. In *Intelligent Virtual Agents*, pages 169–182, 2011.

[23] Hendrik Buschmeier and Stefan Kopp. Understanding how well you understoodcontext-sensitive interpretation of multimodal user feedback. In *Intelligent Virtual Agents*, pages 517–519, 2012.

[24] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jasoslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, Pierre Wellner, and Others. The AMI Meeting Corpus: A Pre-announcement. *Machine Learning for Multimodal Interaction*, 3869:28–39, 2006.

[25] Johanneke Caspers. Melodic characteristics of backchannels in Dutch Map Task dialogues. In *Sixth International Conference on Spoken Language Processing*, 2000.

[26] Justine Cassell, Timothy W. Bickmore, Mark Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pages 520–527, 1999.

[27] Justine Cassell, Alastair J Gill, and Paul A Tepper. Coordination in conversation and rapport. In *Proceedings of the Workshop on Embodied Natural Language*, pages 24–29, 2007.

[28] Nicola Cathcart, Jean Carletta, and Ewan Klein. A shallow model of backchannel continuers in spoken dialogue. *European ACL*, pages 51–58, 2003.

[29] Loredana Cerrato. Some characteristics of feedback expressions in Swedish. In *Proc. of Fonetik*, volume 44, pages 41–44, 2002.

[30] Loredana Cerrato and Mustapha Skhiri. Analysis and measurement of head movements signalling feedback in face-to-face human dialogues. In *Proceedings of the First Nordic Symposium on Multimodal Communication*, pages 43–52, 2003.

[31] Ching-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines.

*ACM Transactions on Intelligent System and Technology*, 2(3):1–27, 2011.

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

[33] Patricia M. Clancy, Sandra A. Thompson, Ryoko Suzuki, and Hongyin Tao. The conversational use of reactive tokens English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3):355–387, 1996.

[34] Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

[35] Herbert H. Clark and Thomas B. Carlson. Hearers and speech acts. *Language*, 58(2):332–373, 1982.

[36] Herbert H. Clark and Meredyth A. Krych. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81, 2004.

[37] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science: A Multidisciplinary Journal*, 13(2):259–294, 1989.

[38] Etienne de Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. Influence of personality traits on backchannel selection. *Intelligent Virtual Agents*, pages 187–193, 2010.

[39] David DeVault, Kenji Sagae, and David Traum. Can I finish?: Learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 11–20. Association for Computational Linguistics, 2009.

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

[41] J.A. Dixon and D.H. Foster. Gender, Social Context and Backchannel Responses. *The Journal of social psychology*, 138(1):134–136, 1998.

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

[43] Kent Drummond and Robert Hopper. Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language and Social Interaction*, 26(2):157–77, 1993.

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

[45] Jens Edlund, Mattias Heldner, and Antoine Pelcé. Prosodic Features of Very Short Utterances in Dialogue. In *Nordic Prosody*, pages 95–106, 2009.

[46] Florian Eyben, Martin Wöllmer, and Björn Schuller. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. 4th International Conference on Affective Computing and Intelligent Interaction*, pages 576–581, 2009.

[47] Marylin S. Feke. Effects of Native Language and Sex on Back Channel Behavior. In *Selected Proceedings from the First Workshop on Spanish Sociolinguistics. Somerville, MA: Cascadilla Proceedings Project*, number 1997, pages 96–106, 2003.

[48] Anna M. Fellegy. Patterns and Functions of Minimal Response. *American Speech*, 70(2):186, 1995.

[49] Jean E. Fox Tree. Listening in on Monologues and Dialogues. *Discourse Processes*,

27(1):35–53, 1999.

[50] Jean E. Fox Tree. Listeners' uses of um and uh in speech comprehension. *Memory & cognition*, 29(2):320–6, March 2001.

[51] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *Proc. Int. Conference on Autonomous Robots and Agents*, pages 379–384, 2004.

[52] Donna T. Fujimoto. Listener Responses in Interaction: A Case for Abandoning the Term, Backchannel. *J Osaka Jogakuin 2 Year Coll*, 37:35–54, 2007.

[53] Rod Gardner. Between speaking and listening: the vocalisation of understandings. *Applied Linguistics*, 19(2):204–224, 1998.

[54] Erving Goffman. *Forms of Talk*. University of Pennsylvania Press, Philadelphia, 1981.

[55] Charles Goodwin. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, pages 97– 121, 1979.

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

[57] Charles Goodwin. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2-3):205–217, 1986.

[58] Marjorie Harness Goodwin. Processes of mutual monitoring implicated in the production of description sequences. *Sociological inquiry*, 50:303– 317, 1980.

[59] Marjorie Harness Goodwin and Charles Goodwin. Concurrent Operations on Talk: Notes on the Interactive Organization of Assessments. *IPRA Papers in Pragmatics*, 1(1):1–54, 1987.

[60] Björn Granström, David House, and Marc Swerts. Multimodal feedback cues in human-machine interactions. In *Speech Prosody*, pages 11–14, 2002.

[61] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, R J Van Der Werf, and Louis-Philippe Morency. Virtual rapport. In *Intelligent Virtual Agents*, pages 14–27, 2006.

[62] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138, 2007.

[63] Agustín Gravano and Julia Hirschberg. Backchannel-Inviting Cues in Task-Oriented Dialogue. In *Interspeech 2009*, pages 1019–1022, 2009.

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

[65] Anna M. Guthrie. On the systematic deployment of okay and mmhmm in academic advising sessions. *Journal of Pragmatics*, 7(3):397–415, 1993.

[66] Uri Hadar, Timothy Steiner, E.C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2(1-2):35–46, 1983.

[67] Uri Hadar, Timothy J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

[68] Reiko Hayashi. Floor structure of English and Japanese conversation. *Journal of Pragmatics*, 16(1):1–30, July 1991.

[69] Bettina Heinz. Backchannel responses as strategic responses in bilingual speakers'

conversations. *Journal of Pragmatics*, 35(7):1113–1142, 2003.

[70] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. Backchannel relevance spaces. In *Proceedings of Nordic Prosody XI*, 2013.

[71] John Heritage. A change-of-state token and aspects of its sequential placement. *Structures of social action: Studies in conversation analysis*, pages 299–345, 1984.

[72] Dirk Heylen. Listening Heads. In *Modeling Communication with robots and virtual agents*, volume 4930 of *Lecture Notes in Artificial Intelligence*, pages 241–259. 2008.

[73] Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. Searching for prototypical facial feedback signals. In *Intelligent Virtual Agents*, pages 147–153, 2007.

[74] Dirk Heylen and Rieks op den Akker. Computing backchannel distributions in multiparty conversations. In *Proceedings of the Workshop on Embodied Language Processing*, pages 17–24, 2007.

[75] Lynette Hirschman. Female-male differences in Conversational Interaction. *Language in Society*, 23:427–442, 1994.

[76] Beth Ann Hockey. Prosody and the role of okay and uh-huh in discourse. *Methods*, pages 128–136, 1993.

[77] Donald Horton and R. Richard Wohl. Mass Communication and Para-social Interaction: Observations on Intimacy at a Distance. *Psychiatry*, 19:215–29, 1956.

[78] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Learning Backchannel Prediction Model from Parasocial Consensus Sampling : A Subjective Evaluation. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 159–172, 2010.

[79] Lixing Huang, Louis-Philippe Morency, and Jonathan 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.

[80] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual Rapport 2.0. In *Intelligent Virtual Agents*, pages 68–79, 2011.

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

[82] Yuri Iwano, Shioya Kageyama, Emi Morikawa, Shu Nakazato, and Katsuhiko Shirai. Analysis of head movements and its role in spoken dialogue. In *Spoken Language Processing*, number Figure 1, pages 2167–2170, 1996.

[83] Gail Jefferson. Notes on a Systematic Deployment of the Acknowledgement Tokens "Yeah" and "Mm hm". In *Tilburg papers in language and literature*, 1983.

[84] Gail Jefferson. Is no an acknowledgment token? Comparing American and British uses of (+)/(-) tokens. *Journal of pragmatics*, 34(10-11):1345–1383, 2002.

[85] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In Oliver P. John, Richard W. Robins, and Lawrence A. Pervin, editors, *Handbook of personality: Theory and research*, chapter 4, pages 114–158. Guilford Press, New York, New York, USA, 3 edition, 2008.

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

[87] Sotaro Kita and Sachiko Ide. Nodding, aizuchi , and final particles in Japanese conversation: How conversation reflects the ideology of communication and social relationships. *Journal of Pragmatics*, 39(7):1242–1254, 2007.

[88] Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seiichi Nakagawa. Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems. *Transactions of the Japanese Society for Artificial Intelligence*, 20:220–228, 2005.

[89] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu 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.

[90] Stefan Kopp, Jens Allwood, Karl Grammer, Elisabeth Ahlsén, and Thorsten Stocksmeier. Modeling Embodied Feedback with Virtual Humans. *Modeling Communication with Robots and Virtual Humans*, pages 18–37, 2008.

[91] Robert M. Krauss, Connie M. Garlock, Peter D. Bricker, and Lee E. McMahon. The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, 35(7):523–529, 1977.

[92] Robert M. Krauss and Sidney Weinheimer. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of personality and social psychology*, 4(3):343–6, 1966.

[93] Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4):718–731, 1982.

[94] Klaus Krippendorff. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

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

[96] Campbell Leaper, Mary Carson, Carilyn Baker, Heithre Holliday, and Sharon Myers. Self-disclosure and listener verbal support in same-gender and cross-gender friends' conversations. *Sex Roles*, 33(5-6):387–404, 1995.

[97] Harold J. Leavitt and Ronald A. H. Mueller. Some Effects of Feedback on Communication. *Human Relations*, 4:401–410, 1951.

[98] Stephen C. Levinson. *Pragmatics*. Cambridge University Press, Cambridge, 1983.

[99] Gina-Anne Levow, Susan Duncan, and Edward T. King. Cross-cultural Investigation of Prosody in Verbal Feedback in Interactional Rapport. In *nterspeech 2010*, pages 286–289, 2010.

[100] Gina-Anne Levow and Siwei Wang. Employing boosting to compare cues to verbal feedback in multi-lingual dialog. *Workshop on Spoken Language Technology*, pages 67–72, 2012.

[101] Kristina Lundholm Fors. The temporal relationship between feedback and pauses: a pilot study. In *Feedback Behaviors in Dialog*, pages 43–45, 2012.

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

[103] Tammy A. Marche and Carole Peterson. On the gender differential use of listener responsiveness. *Sex Roles*, 29(11-12):795–816, 1993.

[104] Senko K. Maynard. Conversation management in contrast: Listener response in Japanese and American English. *Journal of Pragmatics*, 14(3):397–412, June 1990.

[105] Daniel N. McIntosh. Facial feedback hypotheses: Evidence, implications, and directions. *Motivation and Emotion*, 20(2):121–147, 1996.

[106] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Context-based Recognition during Human Interactions: Automatic Feature Selection and Encoding Dictionary. In *10th International Conference on Multimodal Interfaces*, pages 181–188, Chania, Crete, Greece, 2008. ACM.

[107] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *Intelligent Virtual Agents*, pages 176–190, 2008.

[108] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2011.

[109] Anthony Mulac, Karen T. Erlandson, W. Jeffrey Farrar, Jennifer S. Hallett, Jennifer L. Molloy, and Margaret E. Prescott. "Uh-huh. What's That All About?": Differing Interpretations of Conversational Backchannels and Questions as Sources of Miscommunication Across Gender Boundaries. *Communication Research*, 25(6):641–668, 1998.

[110] Daniel Neiberg and Joakim Gustafson. Predicting speaker changes and listener responses with and without eye-contact. *Interspeech*, pages 1565–1568, 2011.

[111] Daniel Neiberg, Giampiero Salvi, and Joakim Gustafson. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55(3):451–469, 2013.

[112] Daniel Neiberg and Khiet P. Truong. Online detection of vocal listener responses with maximum latency constraints. In *Acoustics, Speech and Signal Processing*, pages 7–10, 2011.

[113] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. A spoken dialog system for chat-like conversations considering response timing. In *Text, Speech and Dialogue*, pages 599–606. Springer, 2007.

[114] Hiroaki Noguchi and Yasuharu Den. Prosody-based detection of the context of backchannel responses. In *Spoken Language Processing*, 1998.

[115] 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.

[116] Derya Ozkan and Louis-Philippe Morency. Concensus of Self-Features for Nonverbal Behavior Analysis. In *Human Behavior Understanding*, 2010.

[117] Derya Ozkan and Louis-Philippe Morency. Latent Mixture of Discriminative Experts. *IEEE Transaction on Multimedia*, 15(2):326–338, 2013.

[118] Derya Ozkan, Kenji Sagae, and Louis-Philippe Morency. Latent Mixture of Discriminative Experts for Multimodal Prediction Modeling. In *Computational Linguistics*, pages 860–868, 2010.

[119] Patrizia Paggio and Costanza Navarretta. Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. *Journal on Multimodal User Interfaces*, 7(1-2):29–37, 2012.

[120] Sathish Pammi and Marc Schröder. Evaluating the meaning of synthesized listener

vocalizations. In *Interspeech*, 2011.

[121] Ronald Poppe, Mark ter Maat, and Dirk Heylen. Online Backchannel Synthesis Evaluation with the Switching Wizard of Oz. In *Workshop on Real-Time Conversations with Virtual Agents*, number 1, pages 1–8, 2012.

[122] Ronald Poppe, Khiet Truong, and Dirk Heylen. Backchannels: Quantity, Type and Timing Matters. In *Intelligent Virtual Agents*, pages 228–239, 2011.

[123] Ronald Poppe, Khiet P. Truong, and Dirk Heylen. Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-Agent Systems*, 2013.

[124] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. Backchannel Strategies for Artificial Listeners. In *Intelligent Virtual Agents*, pages 146–158, 2010.

[125] Ken Prepin, Magalie Ochs, and Catherine Pelachaud. Beyond backchannels: co-construction of dyadic stance by reciprocal reinforcement of smiles between virtual agents. In *Proceedings of COGSCI 2013 The Annual Meeting of the Cognitive Science Society*, pages 37–39, 2013.

[126] Julie Reid. A study of gender differences in minimal responses. *Journal of Pragmatics*, 24(5):489–512, 1995.

[127] Jennifer Robison, Scott McQuiggan, and James Lester. Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems. In *Affective Computing and Intelligent Interaction*, pages 1–6, 2009.

[128] Howard M. Rosenfeld. Instrumental affiliative functions of facial and gestural expressions. *Journal of Personality and Social Psychology*, 4(1):65–72, 1966.

[129] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4):696–735, 1974.

[130] Marc Schröder, Dirk Heylen, and Isabella Poggi. Perception of non-verbal emotional listener feedback. *Speech prosody*, pages 1–4, 2006.

[131] Burr Settles. Active learning literature survey. Technical report, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

[132] Candace L. Sidner and Christopher Lee. Attentional gestures in dialogues between people and robots. In Toyoaki Nishida, editor, *Engineering Approaches to Conversational Informatics.*, chapter 6, pages 1–15. 2007.

[133] Gabriel Skantze. A Testbed for Examining the Timing of Feedback using a Map Task. In *Feedback Behaviors in Dialog*, pages 69–72, 2012.

[134] Gabriel Skantze, David House, and Jens Edlund. User Responses to Prosodic Variation in Fragmentary Grounding Utterances in Dialog. In *Ninth International Conference on Spoken Language Processing*, 2006.

[135] Maria Stubbe. Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3):257–289, 1998.

[136] Masashi Takeuchi, Norihide Kitaoka, and Seiichi Nakagawa. Timing detection for real-time dialog systems using prosodic and linguistic information. *International Conference on Speech Prosody*, pages 529–532, 2004.

[137] Peter C. Terry, Andrew M. Lane, and Gerard J. Fogarty. Construct validity of the Profile of Mood States-Adolescents for use with adults. *Psychology of Sport and Exercise*, 4(2):125–139, 2003.

[138] Kristinn R. Thórisson. *Communicative humanoids: a computational model of psychoso-*

*cial dialogue skills*. PhD thesis, Massachusetts Institute of Technology, 1996.

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

[140] David Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella. Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation. In *Intelligent Virtual Agents*, pages 275–288, 2012.

[141] Khiet Truong, Ronald Poppe, Iwan de Kok, and Dirk Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *interspeech*, pages 2973–2976, 2011.

[142] Herwin van Welbergen, Dennis Reidsma, Zsófia M. Ruttkay, and Job Zwiers. Elckerlyc - A BML Realizer for continuous, multimodal interaction with a Virtual Human. *Journal on Multimodal User Interfaces*, 3(4):271–284, 2010.

[143] Astrid von der Pütten, Christian Reipen, Antje Wiedmann, Stefan Kopp, and Nicole C. Krämer. Comparing Emotional vs. Envelope Feedback for Embodied Conversational Agents. In *Intelligent Virtual Agents*, pages 550–551, 2008.

[144] Astrid von der Pütten, Christian Reipen, Antje Wiedmann, Stefan Kopp, and Nicole C. Krämer. The Impact of Different Embodied Agent-Feedback on Users Behavior. In *Intelligent Virtual Agents*, pages 549–551, 2009.

[145] Å sa Wallers, Jens Edlund, and Gabriel Skantze. The effect of prosodic features on the interpretation of synthesised backchannels. In *International Tutorial and Research Workshop*, volume 4021, page 183, Kloster Irsee, Germany, 2006. Springer.

[146] Ning Wang and Jonathan Gratch. Rapport and Facial Expression. In *Affective Computing and Intelligent Interaction (ACII 2009)*, 2009.

[147] Ning Wang and Jonathan Gratch. Don't Just Stare at Me! In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, pages 1241–1249, 2010.

[148] Zhiyang Wang, Jina Lee, and Stacy Marsella. Towards More Comprehensive Listening Behavior: Beyond the Bobble Head. In *Intelligent Virtual Agents*, pages 216–227. Springer, 2011.

[149] Nigel Ward and Yaffa Al Bayyari. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In *Interspeech 2006 Proceedings*, 2006.

[150] Nigel Ward and Yaffa Al Bayyari. A prosodic feature that invites back-channels in Egyptian Arabic. *Perspectives on Arabic Linguistics*, 20:187–206, 2007.

[151] Nigel Ward and Joshua L. McCartney. Visualization to support the discovery of prosodic contours related to turn-taking. Technical Report Figure 1, Tech. Rep. UTEP-CS-10-24, University of El Pao, 2010.

[152] Nigel Ward and Joshua L. McCartney. Visualizations Supporting the Discovery of Prosodic Contours Related to Turn-Taking. In *Feedback Behaviors in Dialog*, pages 85–89, 2012.

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

[154] Tmio Watanabe and Naohiko 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.

[155] David Watson and Lee Anna Clark. The PANAS-X. Technical report, 1994.

[156] Ron White. Back channelling, repair, pausing, and private speech. *Applied Linguistics*, 18(3):314, 1997.

[157] Sheida White. Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, 18(1):59–76, 1989.

[158] James P. Wolf. The effects of backchannels on fluency in L2 oral task production. *System*, 36(2):279–294, 2008.

[159] Deng Xudong. The Use of Listener Responses in Mandarin Chinese and Australian English Conversations. *Pragmatics*, 18(2):303–328, 2010.

[160] Victor H. Yngve. On getting a word in edgewise. In *Sixth Regional Meeting of the Chicago Linguistic Society*, volume 6, pages 657–677, 1970.

[161] Don H. Zimmerman. Acknowledgment Tokens and Speakership Incipiency Revisited. *Research on Language and Social Interaction*, 26(2):179–94, 1993.

# SIKS Dissertations

**2013-28** **Frans van der Sluis (UT),** *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

**2013-27** **Mohammad Huq (UT),** *Inference-based Framework Managing Data Provenance*

**2013-26** **Alireza Zarghami (UT),** *Architectural Support for Dynamic Homecare Service Provisioning*

**2013-25** **Agnieszka Anna Latoszek-Berendsen (UM),** *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

**2013-24** **Haitham Bou Ammar (UM),** *Automated Transfer in Reinforcement Learning*

**2013-23** **Patricio de Alencar Silva(UvT),** *Value Activity Monitoring*

**2013-22** **Tom Claassen (RUN),** *Causal Discovery and Logic*

**2013-21** **Sander Wubben (UvT),** *Text-to-text generation by monolingual machine translation*

**2013-20** **Katja Hofmann (UvA),** *Fast and Reliable Online Learning to Rank for Information Retrieval*

**2013-19** **Renze Steenhuizen (TUD),** *Coordinated Multi-Agent Planning and Scheduling*

**2013-18** **Jeroen Janssens (UvT),** *Outlier Selection and One-Class Classification*

**2013-17** **Koen Kok (VU),** *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

**2013-16** **Eric Kok (UU),** *Exploring the practical benefits of argumentation in multi-agent deliberation*

**2013-15** **Daniel Hennes (UM),** *Multiagent Learning - Dynamic Games and Applications*

**2013-14** **Jafar Tanha (UVA),** *Ensemble Approaches to Semi-Supervised Learning Learning*

**2013-13** **Mohammad Safiri(UT),** *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

**2013-12** **Marian Razavian(VU),** *Knowledge-driven Migration to Services*

**2013-11** **Evangelos Pournaras(TUD),** *Multi-level Reconfigurable Self-organization in Overlay Services*

**2013-10** **Jeewanie Jayasinghe Arachchige(UvT),** *A Unified Modeling Framework for Service Design.*

**2013-09** **Fabio Gori (RUN),** *Metagenomic Data Analysis: Computational Methods and Applications*

**2013-08** **Robbert-Jan Merk(VU),** *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

**2013-07** **Giel van Lankveld (UT),** *Quantifying Individual Player Differences*

**2013-06** **Romulo Goncalves(CWI),** *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

**2013-05** **Dulce Pumareja (UT),** *Groupware Requirements Evolutions Patterns*

**2013-04** **Chetan Yadati(TUD),** *Coordinating autonomous planning and scheduling*

**2013-03** **Szymon Klarman (VU),** *Reasoning with Contexts in Description Logics*

**2013-02 Erietta Liarou (CWI),** *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*
**2013-01 Viorel Milea (EUR),** *News Analytics for Financial Decision Support*
**2012-51 Jeroen de Jong (TUD),** *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*
**2012-50 Steven van Kervel (TUD),** *Ontologogy driven Enterprise Information Systems Engineering*
**2012-49 Michael Kaisers (UM),** *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*
**2012-48 Jorn Bakker (TUE),** *Handling Abrupt Changes in Evolving Time-series Data*
**2012-47 Manos Tsagkias (UVA),** *Mining Social Media: Tracking Content and Predicting Behavior*
**2012-46 Simon Carter (UVA),** *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*
**2012-45 Benedikt Kratz (UvT),** *A Model and Language for Business-aware Transactions*
**2012-44 Anna Tordai (VU),** *On Combining Alignment Techniques*
**2012-43 Withdrawn,**
**2012-42 Dominique Verpoorten (OU),** *Reflection Amplifiers in self-regulated Learning*
**2012-41 Sebastian Kelle (OU),** *Game Design Patterns for Learning*
**2012-40 Agus Gunawan (UvT),** *Information Access for SMEs in Indonesia*
**2012-39 Hassan Fatemi (UT),** *Risk-aware design of value and coordination networks*
**2012-38 Selmar Smit (VU),** *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*
**2012-37 Agnes Nakakawa (RUN),** *A Collaboration Process for Enterprise Architecture Creation*
**2012-36 Denis Ssebugwawo (RUN),** *Analysis and Evaluation of Collaborative Modeling Processes*
**2012-35 Evert Haasdijk (VU),** *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics*
**2012-34 Pavol Jancura (RUN),** *Evolutionary analysis in PPI networks and applications*
**2012-33 Rory Sie (OUN),** *Coalitions in Cooperation Networks (COCOON)*
**2012-32 Wietske Visser (TUD),** *Qualitative multi-criteria preference representation and reasoning*
**2012-31 Emily Bagarukayo (RUN),** *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*
**2012-30 Alina Pommeranz (TUD),** *Designing Human-Centered Systems for Reflective Decision Making*
**2012-29 Almer Tigelaar (UT),** *Peer-to-Peer Information Retrieval*
**2012-28 Nancy Pascall (UvT),** *Engendering Technology Empowering Women*
**2012-27 Hayrettin Gurkok (UT),** *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*
**2012-26 Emile de Maat (UVA),** *Making Sense of Legal Text*
**2012-25 Silja Eckartz (UT),** *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*
**2012-24 Laurens van der Werff (UT),** *Evaluation of Noisy Transcripts for Spoken Document Retrieval*
**2012-23 Christian Muehl (UT),** *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*
**2012-22 Thijs Vis (UvT),** *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*
**2012-21 Roberto Cornacchia (TUD),** *Querying Sparse Matrices for Information Retrieval*

**2012-20 Ali Bahramisharif (RUN),** *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*
**2012-19 Helen Schonenberg (TUE),** *What's Next? Operational Support for Business Process Execution*
**2012-18 Eltjo Poort (VU),** *Improving Solution Architecting Practices*
**2012-17 Amal Elgammal (UvT),** *Towards a Comprehensive Framework for Business Process Compliance*
**2012-16 Fiemke Both (VU),** *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment*
**2012-15 Natalie van der Wal (VU),** *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.*
**2012-14 Evgeny Knutov(TUE),** *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*
**2012-13 Suleman Shahid (UvT),** *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*
**2012-12 Kees van der Sluijs (TUE),** *Model Driven Design and Data Integration in Semantic Web Information Systems*
**2012-11 J.C.B. Rantham Prabhakara (TUE),** *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*
**2012-10 David Smits (TUE),** *Towards a Generic Distributed Adaptive Hypermedia Environment*
**2012-09 Ricardo Neisse (UT),** *Trust and Privacy Management Support for Context-Aware Service Platforms*
**2012-08 Gerben de Vries (UVA),** *Kernel Methods for Vessel Trajectories*
**2012-07 Rianne van Lambalgen (VU),** *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*
**2012-06 Wolfgang Reinhardt (OU),** *Awareness Support for Knowledge Workers in Research Networks*
**2012-05 Marijn Plomp (UU),** *Maturing Interorganisational Information Systems*
**2012-04 Jurriaan Souer (UU),** *Development of Content Management System-based Web Applications*
**2012-03 Adam Vanya (VU),** *Supporting Architecture Evolution by Mining Software Repositories*
**2012-02 Muhammad Umair(VU),** *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*
**2012-01 Terry Kakeeto (UvT),** *Relationship Marketing for SMEs in Uganda*
**2011-49 Andreea Niculescu (UT),** *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*
**2011-48 Mark Ter Maat (UT),** *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*
**2011-47 Azizi Bin Ab Aziz(VU),** *Exploring Computational Models for Intelligent Support of Persons with Depression*
**2011-46 Beibei Hu (TUD),** *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*
**2011-45 Herman Stehouwer (UvT),** *Statistical Language Models for Alternative Sequence Selection*
**2011-44 Boris Reuderink (UT),** *Robust Brain-Computer Interfaces*
**2011-43 Henk van der Schuur (UU),** *Process Improvement through Software Operation Knowledge*
**2011-42 Michal Sindlar (UU),** *Explaining Behavior through Mental State Attribution*
**2011-41 Luan Ibraimi (UT),** *Cryptographically Enforced Distributed Data Access Control*

**2011-40 Viktor Clerc (VU),** *Architectural Knowledge Management in Global Software Development*
**2011-39 Joost Westra (UU),** *Organizing Adaptation using Agents in Serious Games*
**2011-38 Nyree Lemmens (UM),** *Bee-inspired Distributed Optimization*
**2011-37 Adriana Burlutiu (RUN),** *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*
**2011-36 Erik van der Spek (UU),** *Experiments in serious game design: a cognitive approach*
**2011-35 Maaike Harbers (UU),** *Explaining Agent Behavior in Virtual Training*
**2011-34 Paolo Turrini (UU),** *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*
**2011-33 Tom van der Weide (UU),** *Arguing to Motivate Decisions*
**2011-32 Nees-Jan van Eck (EUR),** *Methodological Advances in Bibliometric Mapping of Science*
**2011-31 Ludo Waltman (EUR),** *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*
**2011-30 Egon L. van den Broek (UT),** *Affective Signal Processing (ASP): Unraveling the mystery of emotions*
**2011-29 Faisal Kamiran (TUE),** *Discrimination-aware Classification*
**2011-28 Rianne Kaptein(UVA),** *Effective Focused Retrieval by Exploiting Query Context and Document Structure*
**2011-27 Aniel Bhulai (VU),** *Dynamic website optimization through autonomous management of design patterns*
**2011-26 Matthijs Aart Pontier (VU),** *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*
**2011-25 Syed Waqar ul Qounain Jaffry (VU)),** *Analysis and Validation of Models for Trust Dynamics*
**2011-24 Herwin van Welbergen (UT),** *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*
**2011-23 Wouter Weerkamp (UVA),** *Finding People and their Utterances in Social Media*
**2011-22 Junte Zhang (UVA),** *System Evaluation of Archival Description and Access*
**2011-21 Linda Terlouw (TUD),** *Modularization and Specification of Service-Oriented Systems*
**2011-20 Qing Gu (VU),** *Guiding service-oriented software engineering - A view-based approach*
**2011-19 Ellen Rusman (OU),** *The Mind ' s Eye on Personal Profiles*
**2011-18 Mark Ponsen (UM),** *Strategic Decision-Making in complex games*
**2011-17 Jiyin He (UVA),** *Exploring Topic Structure: Coherence, Diversity and Relatedness*
**2011-16 Maarten Schadd (UM),** *Selective Search in Games of Different Complexity*
**2011-15 Marijn Koolen (UvA),** *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*
**2011-14 Milan Lovric (EUR),** *Behavioral Finance and Agent-Based Artificial Markets*
**2011-13 Xiaoyu Mao (UvT),** *Airport under Control. Multiagent Scheduling for Airport Ground Handling*
**2011-12 Carmen Bratosin (TUE),** *Grid Architecture for Distributed Process Mining*
**2011-11 Dhaval Vyas (UT),** *Designing for Awareness: An Experience-focused HCI Perspective*
**2011-10 Bart Bogaert (UvT),** *Cloud Content Contention*
**2011-09 Tim de Jong (OU),** *Contextualised Mobile Media for Learning*
**2011-08 Nieske Vergunst (UU),** *BDI-based Generation of Robust Task-Oriented Dialogues*
**2011-07 Yujia Cao (UT),** *Multimodal Information Presentation for High Load Human Computer Interaction*
**2011-06 Yiwen Wang (TUE),** *Semantically-Enhanced Recommendations in Cultural Heritage*

**2011-05 Base van der Raadt (VU),** *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.*
**2011-04 Hado van Hasselt (UU),** *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference*
**2011-03 Jan Martijn van der Werf (TUE),** *Compositional Design and Verification of Component-Based Information Systems*
**2011-02 Nick Tinnemeier(UU),** *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*
**2011-01 Botond Cseke (RUN),** *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*
**2010-53 Edgar Meij (UVA),** *Combining Concepts and Language Models for Information Access*
**2010-52 Peter-Paul van Maanen (VU),** *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*
**2010-51 Alia Khairia Amin (CWI),** *Understanding and supporting information seeking tasks in multiple sources*
**2010-50 Bouke Huurnink (UVA),** *Search in Audiovisual Broadcast Archives*
**2010-49 Jahn-Takeshi Saito (UM),** *Solving difficult game positions*
**2010-48 Withdrawn,**
**2010-47 Chen Li (UT),** *Mining Process Model Variants: Challenges, Techniques, Examples*
**2010-46 Vincent Pijpers (VU),** *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*
**2010-45 Vasilios Andrikopoulos (UvT),** *A theory and model for the evolution of software services*
**2010-44 Pieter Bellekens (TUE),** *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*
**2010-43 Peter van Kranenburg (UU),** *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*
**2010-42 Sybren de Kinderen (VU),** *Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach*
**2010-41 Guillaume Chaslot (UM),** *Monte-Carlo Tree Search*
**2010-40 Mark van Assem (VU),** *Converting and Integrating Vocabularies for the Semantic Web*
**2010-39 Ghazanfar Farooq Siddiqui (VU),** *Integrative modeling of emotions in virtual agents*
**2010-38 Dirk Fahland (TUE),** *From Scenarios to components*
**2010-37 Niels Lohmann (TUE),** *Correctness of services and their composition*
**2010-36 Jose Janssen (OU),** *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification*
**2010-35 Dolf Trieschnigg (UT),** *Proof of Concept: Concept-based Biomedical Information Retrieval*
**2010-34 Teduh Dirgahayu (UT),** *Interaction Design in Service Compositions*
**2010-33 Robin Aly (UT),** *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*
**2010-32 Marcel Hiel (UvT),** *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*
**2010-31 Victor de Boer (UVA),** *Ontology Enrichment from Heterogeneous Sources on the Web*
**2010-30 Marieke van Erp (UvT),** *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval*
**2010-29 Stratos Idreos(CWI),** *Database Cracking: Towards Auto-tuning Database Kernels*
**2010-28 Arne Koopman (UU),** *Characteristic Relational Patterns*
**2010-27 Marten Voulon (UL),** *Automatisch contracteren*

**2010-26 Ying Zhang (CWI),** *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*
**2010-25 Zulfiqar Ali Memon (VU),** *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*
**2010-24 Dmytro Tykhonov,** *Designing Generic and Efficient Negotiation Strategies*
**2010-23 Bas Steunebrink (UU),** *The Logical Structure of Emotions*
**2010-22 Michiel Hildebrand (CWI),** *End-user Support for Access to Heterogeneous Linked Data*
**2010-21 Harold van Heerde (UT),** *Privacy-aware data management by means of data degradation*
**2010-20 Ivo Swartjes (UT),** *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*
**2010-19 Henriette Cramer (UvA),** *People's Responses to Autonomous and Adaptive Systems*
**2010-18 Charlotte Gerritsen (VU),** *Caught in the Act: Investigating Crime by Agent-Based Simulation*
**2010-17 Spyros Kotoulas (VU),** *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*
**2010-16 Sicco Verwer (TUD),** *Efficient Identification of Timed Automata, theory and practice*
**2010-15 Lianne Bodenstaff (UT),** *Managing Dependency Relations in Inter-Organizational Models*
**2010-14 Sander van Splunter (VU),** *Automated Web Service Reconfiguration*
**2010-13 Gianluigi Folino (RUN),** *High Performance Data Mining using Bio-inspired techniques*
**2010-12 Susan van den Braak (UU),** *Sensemaking software for crime analysis*
**2010-11 Adriaan Ter Mors (TUD),** *The world according to MARP: Multi-Agent Route Planning*
**2010-10 Rebecca Ong (UL),** *Mobile Communication and Protection of Children*
**2010-09 Hugo Kielman (UL),** *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*
**2010-08 Krzysztof Siewicz (UL),** *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*
**2010-07 Wim Fikkert (UT),** *Gesture interaction at a Distance*
**2010-06 Sander Bakkes (UvT),** *Rapid Adaptation of Video Game AI*
**2010-05 Claudia Hauff (UT),** *Predicting the Effectiveness of Queries and Retrieval Systems*
**2010-04 Olga Kulyk (UT),** *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*
**2010-03 Joost Geurts (CWI),** *A Document Engineering Model and Processing Framework for Multimedia documents*
**2010-02 Ingo Wassink (UT),** *Work flows in Life Science*
**2010-01 Matthijs van Leeuwen (UU),** *Patterns that Matter*
**2009-46 Loredana Afanasiev (UvA),** *Querying XML: Benchmarks and Recursion*
**2009-45 Jilles Vreeken (UU),** *Making Pattern Mining Useful*
**2009-44 Roberto Santana Tapia (UT),** *Assessing Business-IT Alignment in Networked Organizations*
**2009-43 Virginia Nunes Leal Franqueira (UT),** *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*
**2009-42 Toine Bogers (UvT),** *Recommender Systems for Social Bookmarking*
**2009-41 Igor Berezhnyy (UvT),** *Digital Analysis of Paintings*
**2009-40 Stephan Raaijmakers (UvT),** *Multinomial Language Learning: Investigations into the Geometry of Language*

**2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin),** *Service Substitution – A Behavioral Approach Based on Petri Nets*
**2009-38 Riina Vuorikari (OU),** *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*
**2009-37 Hendrik Drachsler (OUN),** *Navigation Support for Learners in Informal Learning Networks*
**2009-36 Marco Kalz (OUN),** *Placement Support for Learners in Learning Networks*
**2009-35 Wouter Koelewijn (UL),** *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*
**2009-34 Inge van de Weerd (UU),** *Advancing in Software Product Management: An Incremental Method Engineering Approach*
**2009-33 Khiet Truong (UT),** *How Does Real Affect Affect Affect Recognition In Speech?*
**2009-32 Rik Farenhorst (VU) and Remco de Boer (VU),** *Architectural Knowledge Management: Supporting Architects and Auditors*
**2009-31 Sofiya Katrenko (UVA),** *A Closer Look at Learning Relations from Text*
**2009-30 Marcin Zukowski (CWI),** *Balancing vectorized query execution with bandwidth-optimized storage*
**2009-29 Stanislav Pokraev (UT),** *Model-Driven Semantic Integration of Service-Oriented Applications*
**2009-28 Sander Evers (UT),** *Sensor Data Management with Probabilistic Models*
**2009-27 Christian Glahn (OU),** *Contextual Support of social Engagement and Reflection on the Web*
**2009-26 Fernando Koch (UU),** *An Agent-Based Model for the Development of Intelligent Mobile Services*
**2009-25 Alex van Ballegooij (CWI),** *"RAM: Array Database Management through Relational Mapping"*
**2009-24 Annerieke Heuvelink (VUA),** *Cognitive Models for Training Simulations*
**2009-23 Peter Hofgesang (VU),** *Modelling Web Usage in a Changing Environment*
**2009-22 Pavel Serdyukov (UT),** *Search For Expertise: Going beyond direct evidence*
**2009-21 Stijn Vanderlooy (UM),** *Ranking and Reliable Classification*
**2009-20 Bob van der Vecht (UU),** *Adjustable Autonomy: Controling Influences on Decision Making*
**2009-19 Valentin Robu (CWI),** *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*
**2009-18 Fabian Groffen (CWI),** *Armada, An Evolving Database System*
**2009-17 Laurens van der Maaten (UvT),** *Feature Extraction from Visual Data*
**2009-16 Fritz Reul (UvT),** *New Architectures in Computer Chess*
**2009-15 Rinke Hoekstra (UVA),** *Ontology Representation - Design Patterns and Ontologies that Make Sense*
**2009-14 Maksym Korotkiy (VU),** *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*
**2009-13 Steven de Jong (UM),** *Fairness in Multi-Agent Systems*
**2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin),** *Operating Guidelines for Services*
**2009-11 Alexander Boer (UVA),** *Legal Theory, Sources of Law & the Semantic Web*
**2009-10 Jan Wielemaker (UVA),** *Logic programming for knowledge-intensive interactive applications*
**2009-09 Benjamin Kanagwa (RUN),** *Design, Discovery and Construction of Service-oriented Systems*
**2009-08 Volker Nannen (VU),** *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

**2009-07 Ronald Poppe (UT),** *Discriminative Vision-Based Recovery and Recognition of Human Motion*
**2009-06 Muhammad Subianto (UU),** *Understanding Classification*
**2009-05 Sietse Overbeek (RUN),** *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*
**2009-04 Josephine Nabukenya (RUN),** *Improving the Quality of Organisational Policy Making using Collaboration Engineering*
**2009-03 Hans Stol (UvT),** *A Framework for Evidence-based Policy Making Using IT*
**2009-02 Willem Robert van Hage (VU),** *Evaluating Ontology-Alignment Techniques*
**2009-01 Rasa Jurgelenaite (RUN),** *Symmetric Causal Independence Models*