Towards Building an Attentive Artificial Listener

On the Perception of Attentiveness in Audio-Visual Feedback Tokens

Catharine Oertel
KTH
Royal Institute of Technology
Stockholm, Sweden
catha@kth.se

José Lopes
KTH
Royal Institute of Technology
Stockholm, Sweden
jdlopes@kth.se

Yu Yu
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
yu.yu@idiap.ch

Kenneth A. Funes Mora
Idiap Research Institute
Switzerland
kfunes@idiap.ch

Joakim Gustafson
KTH
Royal Institute of Technology
Stockholm, Sweden
jocke@speech.kth.se

Alan W. Black
Carnegie Mellon University
Pittsburgh, United States
awb@cs.cmu.edu

Jean-Marc Odobez
Idiap Research Institute
École Polytechnique Fédérale
de Lausanne (EPFL)
Switzerland
odobez@idiap.ch

ABSTRACT

Current dialogue systems typically lack a variation of audio-visual feedback tokens. Either they do not encompass feedback tokens at all, or only support a limited set of stereotypical functions. However, this does not mirror the subtleties of spontaneous conversations. If we want to be able to build an artificial listener, as a first step towards building an empathetic artificial agent, we also need to be able to synthesize more subtle audio-visual feedback tokens. In this study, we devised an array of monomodal and multimodal binary comparison perception tests and experiments to understand how different realisations of verbal and visual feedback tokens influence third-party perception of the degree of attentiveness. This allowed us to investigate i) which features (amplitude, frequency, duration...) of the visual feedback influences attentiveness perception; ii) whether visual or verbal backchannels are perceived to be more attentive iii) whether the fusion of unimodal tokens with low perceived attentiveness increases the degree of perceived attentiveness compared to unimodal tokens with high perceived attentiveness taken alone; iv) the automatic ranking of audio-visual feedback token in terms of conveyed degree of attentiveness.

ICMI'16, November 12–16, 2016, Tokyo, Japan

http://dx.doi.org/10.1145/2993148.2993188

CCS Concepts

•**Human-centered computing** → **Gestural input;** *Auditory feedback;* Empirical studies in collaborative and social computing;

Keywords

head nods; backchannels; virtual agent

1. INTRODUCTION

Conversational systems are becoming more and more a part of our every–day life. We are opening the phone to ask Siri what the weather will be like tomorrow or switch on our in-car navigational system to tell us how to get from Aachen to Berlin. What most of these systems have in common is their design for fulfilling a task; they are not designed for social interaction. Having a conversational system however, which is able to engage in a social interaction with a human, could open the door to many new applications. For instance, in the educational sector such an application could be the development of a study peer. In healthcare, this could be a virtual therapy aider; a system soldiers suffering from post traumatic stress disorder could talk to and which would not be judgmental.

Requirements for such a system are manifolds. Apart from the understanding and synthesis of content, the system needs also to be able to understand and synthesise the paralinguistic components. One paralinguistic component which is very important in our every–day interactions is *"attention"*. We are constantly searching our interlocutors for their reactions and their level of attention. If we feel that their level of attention is not very high, we will change our behaviour accordingly. However, we also expect our interlocutors' level of attention to fluctuate to a certain degree.

Figure 1: Perception Test. An avatar is displayed and listens to a carrier audio sentence. Given two realizations differing by their backchannels (using resynthesized head nods and inserted audio backchannels from real data), raters are asked to indicate which one looks more attentive, if any.

We would find it very odd if a person would never vary in the degree he or she is paying attention.

The variation in the level of attention in the state of the listener is exactly where current dialogue systems are still limited, however. They are limited in that, if they are including feedback token at all, these are typically representing only a very limited set. These feedback token are generally based on a few recordings of a voice talent, who was told to portray a given feedback function or emotion. It is often fair to assume that these feedback token represent extremes or stereotypes in terms of backchannel realisations. However, the problem with only using such stereotypical backchannels in a dialogue systems is that this is in fact rather atypical of human-human conversations. In reality, people alter nuances in their realisation of backchannels to convey their reaction to the speaker's speech rather than resorting to extremes. In other words, getting bored and consequently disengaging from the conversation is not an instantaneous event but rather a gradual process.

In the current paper we study the degree of attentiveness conveyed by visual (nods), audio, and audio-visual backchannels. The aim is to work towards enabling a dialogue system to portray a listener's level of attentiveness in a more human-like manner. In order to accomplish this, it is essential to have a good understanding of, firstly, which features are important for the perception of different degrees of attentiveness in feedback token and, secondly, to understand how attentiveness is expressed multimodally. For example, should a dialogue system choose to give feedback by using a headnod or should it use an audio-backchannel?

To this end, we rely on a set of perceptual experiments, whose principles are illustrated in Fig. 1, and ultimately propose different models for ranking audio and visual feedback token in terms of their attentiveness. The research questions we aim at answering are:

- **RQ1**: Can third party observers differentiate headnods in terms of attentiveness?
- **RQ2**: If yes, is it possible to identify visual cues which can help quantify these differences?
- **RQ3**: Are headnods or audio backchannels perceived to convey more attentiveness?
- **RQ4**: Does a fusion of headnods and audio backchannels which are ranked low in terms of their perceived attentiveness, increase the perceived level of attentiveness in comparison to high attentively ranked unimodal backchannels.

The rest of the paper is organized as follows: Section II provides background material and our contributions. Section III introduces the backchannels used for the perceptual experiments, which are described in Section IV. Section V presents the results and discussions, while Section 7 concludes the work.

2. BACKGROUND

Backchannels have been studied from both the visual, audio, and multimodal perspective. We review these works below and then provide our contributions.

2.1 Head Nods

As summarized by Malisz et al. in their overview paper [24], there are only few studies over the last decades which have been concerned with the classification and description of head nods. From a backchannel perspective, the most relevant one was carried out by Rosenfeld and Hadris [19] who suggested that smaller, single nods often function as typical backchannel. Moreover, more recently, according to Bousmalis et al. [4], large amplitude repeated nods were found to be characteristic of affirmative meanings. And in [11] Malisz et al. found that distracted listeners, while also generally preferring simple gestures produce complex gestures relatively more frequently than attentive listeners.

2.2 Audio Backchannels

There are only a few studies that investigated the prosodic realisation of backchannel functions such as for instance [17, 12, 10]. Neiberg et al. [12] investigated how prosodic realization influences the perceived function of feedback token taken from dyadic conversations. They found that feedback token often were multi–functional, and some conveyed both understanding, agreement, certainty, and negative surprise. The perceived functions were found to be correlated with prosodic cues. For instance, token with a fast speaking rate and a moderate F0 rise were found to convey understanding and interest. A further study on the prosodic characteristics of feedback expressions was done by Malisz et al. [10]. They analysed the prosodic characteristics of ("ja", "m" and "mhm") across their pragmatic functions, as well as the differences in feedback produced by distracted vs. attentive listeners. By decomposing and analysing feedback signals, they investigated feedback function differences and distractedness-related differences. They found that attentive listeners tend to speak more loudly, energy is less variable, and pitch variability measures are positively related to attentiveness. They argue that prosodic features may strongly depend on segmental structure e.g. nasality vs. orality syllabic structure vs monosyllabic structure.

2.3 Multimodaly and Backchannel

One of the few studies which has focused on exploring which feedback token should be used at which point in time is the work of Poppe et al. [18]. They showed that not only the timing of feedback has an effect on the perception of third party observers, but also the modality of the backchannel. In particular, they investigated whether a backchannel (one head nod, or one audio backchannel, "continuer"), fits better the context of a given sentence.

Truong et al. [23], in another related study, which focused on analysing backchannel (BC) inviting cues in a face-to-face setting, did not find rising/falling pitch to be a BC-inviting cue, as is typical for telephone-style dialogs. In con-

trast, they found gaze to play a more important role, in that mutual gaze occurs significantly more often during visual backchannels. They found also that vocal backchannels are more likely to be timed during pauses in the speaker's speech.Bevacqua et al. [2] investigated the use of multimodal backchannels in conversational agents. More precisely, they focused on the meaning (like agreement, acceptance, refusal) carried by monomodal and multimodal feedback tokens. They found that combined multimodal tokens of the same meaning were not always conveying a stronger attribution of that meaning. They also deduced that the meaning of the multimodal backchannel could not always by simply deduced from the meanings of the components. Finally they found that fused backchannels, composed of backchannels with strong opposite meanings, were rated by third-party observers as being nonsense.

2.4 Backchannels in Virtual Agents and Dialogue Systems

The importance of using backchannels in artificial agents and robots, has been shown in [5, 20, 8, 22, 25]. For instance, [20] investigated the effect of head nod in human robot interactions (HRI). They found that informing subjects that the robot is able to recognize their head nods as well as having the robot provide gestural feedback of its nod recognition is an effective strategy in triggering further nods in the subjects. Other examples include [8], who uses backchannels as one strategy for establishing rapport between a human and a virtual human, or [22], who used backchannels in their dialogue system to study turn-taking in multi-party human robot interactions.

2.5 Paper Contributions

The first contribution of this paper is an in depth unimodal analysis of feedback tokens in terms of their perceived attentiveness. To our knowledge, [11] is the only study which investigated this issue, but from an attentiveness and distractedness perspective. In this work, authors distract a subject and compare the difference in realisation of feedback token across conditions. A first person experience of distractedness is however very different from a third party observer's perception of attentiveness, and therefore the results of the current study and [11] are only conditionally comparable. Moreover, the focus of their paper was a more global analysis of feedback token in terms of distractedness/attentiveness also in relation to various feedback functions. In the current paper, however, we focus on just one kind of audio and visual feedback token (agreement conveying), but therefore analyse it in depth. In addition, one important difference is that in the current study the stimuli are being realised by an avatar, rather than by replaying videos of human–human interactions. This allows to eliminate some other variables (esp. appearance, visual expression) which may affect the perception.

The second contribution is the comparative analysis of visual and audio backchannels and their ranking in terms of attentiveness. To our knowledge, this has not been attempted before.

A third contribution are the different attentiveness ranking models. Not only can they be used to predict the attentiveness of listeners, but they can also serve to rate natural feedbacks like nods extracted from real conversations (thus allowing to exploit a large diversity of feedback realisation) and use them in attentiveness unit–selection feedback synthesisers.

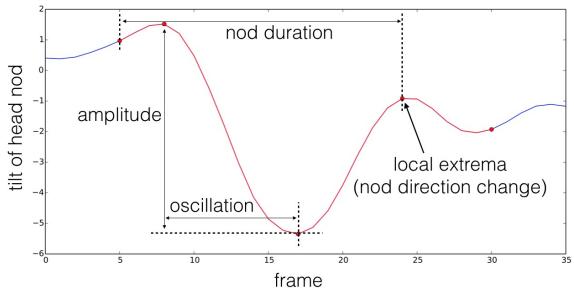

Figure 2: Tilt angle sequence comprising a nod, with a visual representation of the features representing its dynamics.

3. DATA

For this paper, we decided to use natural realisations of audio and visual backchannels to synthesize the avatar's listening behaviors, as shown in Fig. 1. To this end, we used the multi-party group discussion KTH-Idiap corpus [15], in which three PhD students have to convince a Post-Doc that they were the best suited candidate for a prestigious scholarship. We chose this corpus as it is rich in conversational dynamics, especially in terms of participants' degree of attentiveness.

3.1 Audio Backchannels

The set of audio backchannels used in the experiments were selected according to [16]. More precisely, to determine syllable boundaries, syllables were manually annotated in Praat [3]. We then calculated the syllable duration, in milliseconds, from the corresponding TextGrid file and, in a next step, extracted pitch and intensity values using Praat [3]. The selected set contained 64 unique bisyllabic backchannels token sampled across 9 speakers (5 male, 4 female).

3.2 Head Nods

Head nods samples have been selected from the corpus amongst those automatically detected using the method of [6], which relies on a 3D head fitting and pose tracking. The frame rate of the recordings was 30 frames per second. Detected nods have been validated, and those accompanied with speech have been removed. This resulted in 77 nod samples. For these samples, several features characterizing the nod have been extracted, mainly from the head tilt sequence, as shown in Fig. 2.

- **NodDur**: nod duration (in number of frames).
- **NbOsc**: number of oscillations, where an oscillation is defined as the segment between two extremas of the smoothed tilt sequence.
- **NodFreq**: frequency of the nod, computed as $\frac{NOsc}{2*NDur}$.
- **FiDir**: first oscillation direction (up or downward).
- **FiAmp**: absolute amplitude of this oscillation.
- **MaxDownAmp and MaxUpAmp**: maximum amplitude of the downward (resp upward) oscillations.
- **MaxDownSpeed and MaxUpSpeed**: maximum downward (resp. upward) speed.
- **MeanPan**: absolute value of the mean pan (head left-right rotation) during the first oscillation.

Head nods synthesis. The extracted head pose sequences contained some jittering, which, when resynthesised, looked unnatural. To handle this, the six degrees of freedom of the pose were smoothed using a simple moving average filter. A 7-frame smoothing was applied, except for the tilt where shorter windows were used to avoid the elimination of shorter amplitude head nods.

4. PERCEPTION TEST SETUP

We first describe below the overall experimental approach. We then describe the different experiments we have conducted.

4.1 Overall Methodology

Perception Stimuli: Experiments were conducted using the Crowdflower crowdsourcing platform. It consisted in comparisons of the attentiveness level of two feedback realisations, as shown in Fig. 1. In each case, a carrier sentence from a real interaction was used, and backchannel data (audio sample, pose sequence of nods, or combination of these, see Section 3) were inserted at the same place in which a backchannel had occurred in the original recording, and sent into the IRISTK [21] toolkit to realize the virtual agent's listening behavior. Carrier sentences were selected to encompass a backchannel in the original recordings, and be short in duration so as to ensure that the raters could still remember the previous backchannel/head nod when doing comparisons. We emphasize again here that each item of a pair of feedback realizations to be compared was embedded into the same carrier sentence so that it was possible to ensure that backchannel tokens were rated in the same interactional environment.

Ratings: Raters were recruited from the United States, Netherlands, and Germany. They were instructed to watch and listen the video pairs and determine in which video the listener avatar sounded more attentive. An attentive listener had been described to raters as someone who a) pays attention; listens carefully; is observant; b) is careful to fulfill the needs or wants of the speaker; is considerate about the speaker. In a drop–down menu, raters could indicate in which video they perceived the listener to be more attentive or when they could not see any difference. Also, they could report if the video files did not play correctly in their browser. To ensure that we received the best quality ratings, we chose a minimum time threshold of 160 seconds to complete 10 ratings. If a rater was under this threshold (which was based on the average annotation speed of one of the authors), he was automatically discarded. Moreover, we set a maximum of 20 judgments per rater so as to avoid any tiredness effects. Furthermore, we chose the crowdflower settings as to prefer raters with high quality records. Each pair of videos to be compared were annotated by 12 raters.

4.2 Experiment 1 - Head Nods

We were first interested into comparing the attentiveness of different head nod realisations. Pairs of nods to be compared were exhaustively obtained from the 77 unique nods (see Section 3) obtained from 15 persons. Since we wanted to be consistent with prior experiments with audio-only backchannels [16], pairs were only created from nods from the same person, resulting in 439 unique pairs and 5268 ratings.

4.3 Experiment 2 - Nods vs Backchannels

We then wanted to compare the perceived degree of attentiveness of nods versus audio backchannels. Since an exhaustive comparison was not possible, we adopted the following approach. The 5 nods which were rated as most attentive from the ratings according to the methodology described in Section 4.5, as well as the 5 rated the least attentive, were selected. The same was done with the audio backchannels.

From these samples, an exhaustive comparison of nods vs. audio backchannels was conducted, as depicted in Fig. 3. This comprised the 5 top nods vs. the 5 top backchannels, the lowest nods vs.the lowest backchannels, and top vs low audio and visual token comparison (top nods. vs low backchannels, and low nods vs. top backchannels).

4.4 Experiment 3: Bimodal Backchannels

To test whether a bimodal backchannel (a backchannel and a headnod produced at the same time) is perceived to be more attentive than a unimodal backchannel, we adopted the following design. We fused head nods and backchannels taken from the 3 lowest rankings (9 samples in total), and compared them to unimodal backchannels (nods or audio) which were ranked higher than them, as depicted in Fig. 3. We were interested to find whether the resulting bimodal backchannels were perceived to be more attentive than their monomodal components and if yes, to what degree. We hypothesised that the fusion of modalities should increase the perception of the degree of attentiveness. Therefore, we put the bimodal backchannels deliberately at a disadvantage; if bimodal backchannels combined out of low ranking unimodal backchannels are already perceived as more attentive than unimodal ones, then a fusion of higher ranking unimodal backchannel token should be perceived as even more attentive.

4.5 Top-Lesser Tokens Selection

Our experimental approach required to identify within the set of tokens $X = \{\mathbf{x}_i\}_{i=1}^{N}$ the subset that was annotated as more or less attentive from the pairwise crowdsourcing comparisons. Let $\Lambda = \{\mathbf{x}_i, \mathbf{x}_j, v_i, v_j\}_{i,j}$ be the pairwise comparisons, where, v_i and v_j indicate the amount of votes (more attentive) received by the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. To sort the tokens in terms of attentiveness, we take a simple approach:

$$A(\mathbf{x}_i) = \sum_j h(v_i > v_j), \tag{1}$$

where, $A(\mathbf{x}_i)$ is our attentiveness measure and h is a function returning 1 for values equal or above zero, and returning 0 otherwise. In other words, we count the amount of times $\mathbf{x}_i$ is labelled as more attentive.

Nevertheless, to take into account that not all samples were compared to each other, before evaluating Eq. 1, we generated virtual comparisons for the missing pairs as follows: we first defined a directed graph $G_A = \{X, \Omega\}$, where the samples (X) are the nodes in the graph and $\Omega = \{\omega_{i,j}\}_{\{i,j\}}$ are the set of directed edges, whose weight from node i to j is equal to $\omega_{i,j} = v_i - v_j$. For a given pair i, j whose comparison was not made, we first find the shortest path $P_{i,j}(\hat{\Omega})$ (the path with the lowest cumulative weight) in the equivalent graph $\hat{G}_A = \{X, \hat{\Omega}\}$, whose weights $\hat{\Omega}$ are equal to 1 if the comparison was made, and ∞, if not. The weight then assigned to the given pair is equal to the cumulative

weight across the path: $\omega_{i,j} = \sum_{l,k \in P_{i,j}} \omega_{l,k}$. Although this approach is an approximation, we believe this to be sufficient for our purposes, e.g. to identify the top 5 tokens in terms of attentiveness.

5. RESULTS

In this Section we report results from the perception experiments described in the previous section. Here, a sample is said to *win* a comparison if it is labeled as more attentive than the other "losing" sample by 2/3 of the raters. If no 2/3 majority is obtained, the comparison is said to end up in a *tie*.

5.1 Head Nods Analysis

Given the results from Section 4.2 we here report the impact of head nod dynamics, embedded by the features described in Section 3.2, over attentiveness. To this end, we defined two groups of head nods: "Wi" (Winners) and "Lo" (Losers). Then, for each pairwise comparison, the winning head nod (higher rated), is assigned to the "Wi" group, whereas the other sample is assigned to the "Lo" group. We excluded all the head nod pairs for which no 2/3 majority on preference was observed. Note that samples can be repeated within the previously mentioned groups.

The results across the head nods Wi/Lo groups over head nod features can be observed in Figure 4.

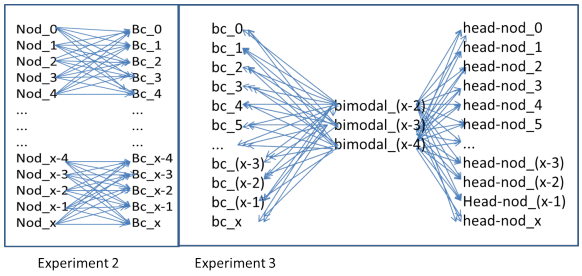

Figure 3: A blue line represents a comparison of, for Experiment 2: (cf. Sec. 4.3)a given head nod and audio backchannel. Experiment 3 :(cf. Sec. 4.4). A fused head nod and audio backchannel (bimodal token) and a monomodal token (head nod or audio backchannel).

Table 1: The relationship between "Wi" and "Lo" and Head Nod features (ns= non significant).

	Winning	Losing	p-vlaue
MaxUpAmp	–	–	ns
MaxDownAmp	M=4.66 SD=2.54	M=3.26 SD=2.60	<0.001
MaxUpSpeed	–	–	ns
MaxDownSpeed	M= 1.17 SD=0.46	M= 0.92 SD=0.51	<0.001
NodDur	M=23.50 SD=16.39	M=16.17 SD=7.16	<0.001
NbOsc	M=4.30 SD=3.57	M=2.94 SD=1.44	<0.001
MeanPan	–	–	ns
FiAmp	M=4.48 SD=3.31	M=3.71 SD=3.14	<0.001

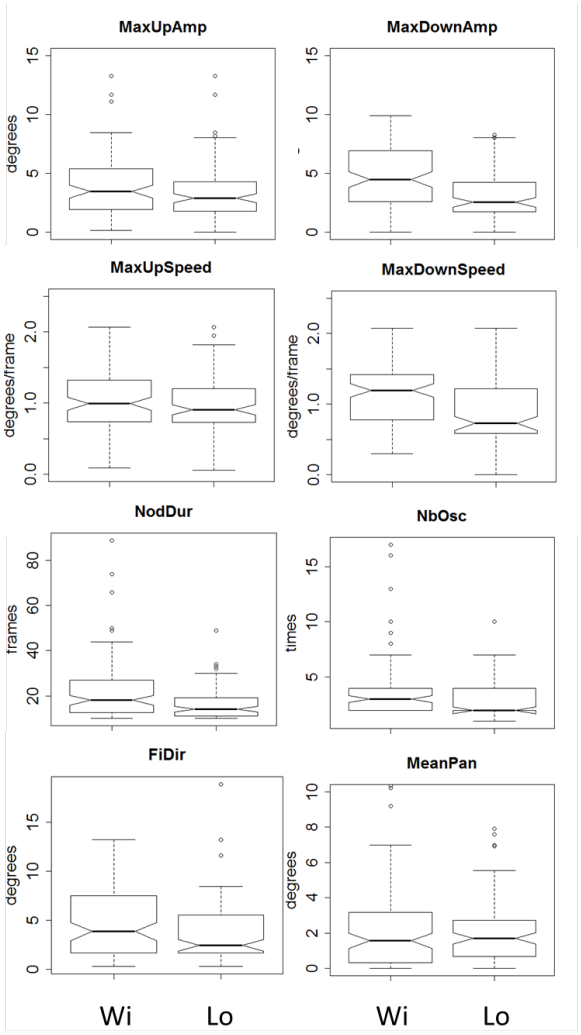

Figure 4: Head nod features distributions within the Winning (Wi) and Losing (Lo) sets.

5.1.1 *Statistical Analysis on Head Nods Features*

Two Sample t-test were conducted to compare NodDur, MaxUpAmp, MaxDownAmp, MaxUpSpeed and MaxDownSpeed, NbOsc, FiAmp and MeanPan between "Wi" and "Lo" stimuli. For these evaluations, 104 pairwise comparisons were available with 2/3 majority.

Results are depicted in Table 1. It can be summarised that the "Wi" head nods have a higher MaxDownAmp, a quicker MaxDownSpeed, a longer NodDur a higher average number of NbOsc as well as a higher FiAmp.

5.2 Head Nod vs. Backchannel Analysis

From the results obtained from the perception experiment described in Section 4.3, a chi-square test was performed to examine the relation between the head nods and audio backchannel modalities and perceived attentiveness. The relation between these variables was significant, X2 (1, N = 90) =88.2 , p <.01. Head nods were more likely to be perceived as more attentive than backchannels.

Further analysis revealed that, out of the 13 comparisons in which the backchannels won over the head nods, 11 contained backchannels ranked in the top 5 of all investigated

backchannels, and only 2 were ranked in the bottom 5. Similarly, out of the 77 comparisons in which head nods won over backchannels, 53 contained head nods ranked in the bottom 5 and 24 in the top 5. These results are depicted in Figure 5. Finally, out of a total number of 200 comparisons between the modalities, 110 ended with a tie.

5.3 Multimodal-Monomodal Comparison

We here present the results for the experiment described on Section 4.4. In order to avoid increasing the number of comparisons times nine, we restricted the analysis to the comparison of two backchannels giver/ listeners. In order to balance for gender, we chose a male and a female participant. Both results are depicted in Figure 6. The results for the female participant will be presented first and will then be followed by the results for the male participant.

A chi-square test was performed to examine the relation between monomodal to multimodal backchannels in terms of perceived attentiveness. To this end, the fused tokens described in Section 4.4 (multimodal samples combining a low ranking backchannel and a low ranking head nod) were compared against two sets of monomodal backchannels: a mid ranked and a top ranked set. The top ranked set consisted of samples labeled with high attentiveness from the pairwise human perception experiments whereas the mid ranked set consisted of samples labeled in the middle of the scale of attentiveness according to the pairwise human perception comparisons. Nevertheless, in both cases, the top five highest ranked samples were not used. The same protocol was repeated comparing the fused tokens vs. a set of monomodal head nods distributed in similar mid/top ranked sets.

For the female participant, the relationship between these variables, when comparing against the mid ranked set, was significant for both nods X2 (1, N = 81) = 35.65 , p <0.001 as well as backchannels X2 (1, N = 81) = 68.48 , p < 0.01. In both cases, the multimodal backchannels won over the monomodal backchannels significantly more often. For the case of comparing against the top ranked set of monomodal samples, even if the obtained difference was not as pronounced, a similar effect could be observed for head nods X2 (1, N =45) = 8.7111, p <0.001. However, the difference to the top ranked set of monomodal audio backchannels, was not significant.

For the male participant, when comparing against the mid ranked set, the difference in terms of number of wins and loses for nods X2 (1, N =45) = 21.511 , p <0.001 as well as backchannel was significant X2 (1, N = 54) = 50.704 , p <0.001. In both cases, the multimodal backchannels won over the monomodal backchannels significantly more often. While the same trend can be observed for the comparison to the top-ranked monomodal backchannels, the trend is not as pronounced for nods X2 (1, N = 45)= 11.378, p <0.001 and for audio backchannel X2 (1, N = 45)=4.4444, p <0.05.

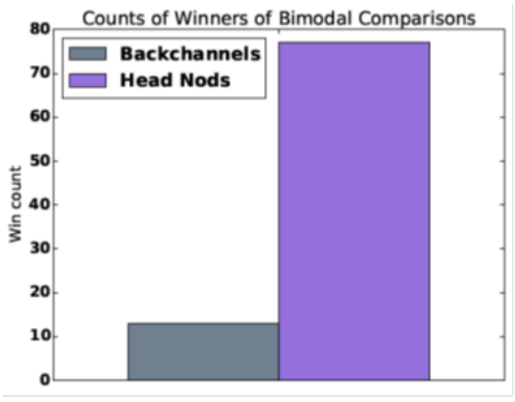

Figure 5: Winning counts obtained for the comparisons between the head nods and backchannel modalities.

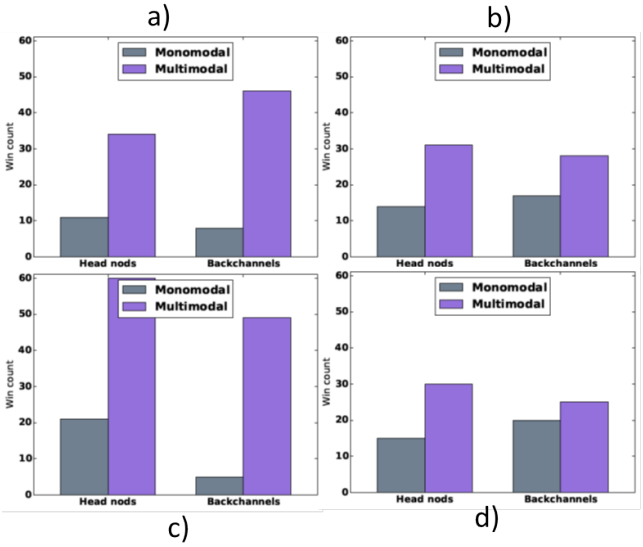

Figure 6: Winning counts for comparisons between multimodal samples and a) mid ranked unimodal samples (male participant); b) top ranked unimodal samples (male participant); c) mid ranked unimodal samples (female participant) and; d)top ranked unimodal samples (female participant)

5.4 RankSVM Classification

In this Section, we evaluated whether it is possible to obtain an automatic assessment on attentiveness. In order to avoid defining an explicit attentiveness scale and stating it as a regression problem, we instead formulated this task as a ranking problem, where two samples are compared according to relative attentiveness. To this end, we employed a Ranking SVM algorithm, in which, for a given pair $\mathbf{x}_i$ and $\mathbf{x}_j$, their difference $(\mathbf{x}_i - \mathbf{x}_j)$ is classified into +1 (if i is more attentive than j) or -1 (if j is more attentive than i), thus turning the problem into a binary classification. To assign such target label (relative attentiveness) to a given pairwise comparison, we used the majority vote from the crowdsourcing experiments.In our experiments we used a Support Vector Machine (SVM) classifier based on the linear kernel. We applied a grid-search with 10-fold cross validation to identify the hyper-parameters.

5.4.1 RankSVM Headnod Results

All cases in which no two-third majority vote by the raters for a comparison was reached, were excluded from the analysis. This reduced the set to 101 tokens. For the classification we used all features which were previously determined to have a significant effect on perceived attentiveness

namely "maximum downward amplitude", "frame duration", "number of oscillations", and first amplitude. As a result of unimodal head nod experiment, we obtained an accuracy of 74.3%.

5.4.2 RankSVM Backchannel Results

To compare to the head nod results of the current paper, we also briefly state the rank SVM results for the backchannels as reported in [16]. All cases in which the majority of raters indicated that there was no perceivable difference in attentiveness, was excluded from the data set. This reduced the size of the data set to 205 pairs. For classification we used all features which were previously determined to have a significant effect on perceived attentiveness except for f0–slope, as this feature reduced the performance of the classifier. The remaining features thus were rms–intensity of first and second syllable as well as the duration of the second syllable. As a result of the unimodal backchannel experiment, we obtained an accuracy of 83%

6. DISCUSSION

6.1 Head Nods

This paper investigated the perception of attentiveness in an array of mono-, bi-, as well as multimodal perception tests. We also investigated which visual features are related to the perception of attentiveness, and investigated whether automatic ranking of visual feedback token was possible. We found that head nods which were perceived as more attentive had a larger maximum downward amplitude. This finding could in part be explained by the findings of [4] et al. who found that a large amplitude is characteristic of affirmative meanings. If someone is perceived to be affirmative it might be reasonable that he is also perceived to be attentive. However, our findings contrast the findings by [11] who found that attentive listeners produce complex head-gestures less frequently than distracted listeners. This difference in findings, however, might simply be due to the difference in study design. While [11] investigate the effects of distracting a participant on his/her realisations of headnods, we investigate the third-party observer perception of attentiveness on different head nod realisations in pairwise comparisons. The kind of nods which [19] describe as typical backchannels seem in our study to be best characterised by the group of backchannels which is perceived to convey less attentiveness. We could not find further studies which investigated perceived attentiveness in headnods. However, for audio backchannels we know that the rms-intensity(loudness) as well as duration are important cues for perceived attentiveness. The louder and the longer the audio backchannel, the more attentive the backchannel is perceived. One might argue that the amplitude is the visual equivalent of the rms-intensity and frame duration and number of oscillations are the visual equivalents of duration. Therefore, the fact that particularly these features have a significant effect on the perceived attentiveness in head nods might make sense when observed from a multimodal perspective.

6.2 Bimodal Comparisons

The fact that nearly all backchannels lost the comparisons to the head nods might partly explained by diverse findings from the literature: For example, [12] found that tokens with a fast speaking rate and a moderate F0-rise were found to convey understanding and interest; also, [9] came to a similar conclusion that a high or rising pitch is perceived as more engaged than a flat pitch. In fact, however, this kind of backchannel is very rare or even not existent in the backchannels produced by the speakers in this corpus. While we found significant differences in the perception of attentiveness, in audio backchannels in this corpus, these differences were found in duration, rms-intensity, and F0-slope of the first syllable.

Therefore, it might be argued that the audio backchannels of this corpus are situated in the lower ends of attentiveness and the participants in the corpus recordings chose to express their attentiveness maybe by means of visual rather than verbal means. Moreover, the timing of backchannels plays of course an important role. We tried to control for it by using the backchannels' timings of the original corpus. In case of the bimodal comparisons dicussed here, we used the timings of the audio backchannel which potentially might have put the visual backchannel at a disadvantage. However, results show that despite this disadvantage, the visual backchannels perform better. This can either indicate that the visual modality is the more dominant one in terms of perceived attentiveness or/and that the speaker sentences were too short for the different timings to play a role.

6.3 Multimodal-Monomodal Comparison

The study design of the current multimodal-monomodal comparison experiment is, in many ways, comparable to Bevacqua et al.'s study [2]. Similar to [2], we also investigated the effect the fusion of monomodal backchannels into multimodal backchannel tokens has on the perception of third-party observers. While Bevacqua et al. focused on the conveyance of a given meaning, however, we focused on the perception of attentiveness. The authors found that the strongest attribution of meaning is not always the multimodal one. In the current paper the fusion of modalities similarily did not always lead to the highest perception of attentiveness. However, except for one comparison condition, fused backchannels vs. the top-five-ranked monomodal audio backchannels, the number of times the multimodal backchannel won over the monomodal backchannel was significant. It has to be noted as well, that the comparisons were biased towards favouring the mono-modal audio-visual back-channel. We only fused the backchannel tokens which were perceived lowest in terms of attentiveness and compared them to higher ranked mono-modal tokens; if we had fused higher ranked backchannel tokens, the multimodal backchannel tokens might have won even more often over the mono-modal ones.

7. CONCLUSION

In this paper, we have shown that "maximum downward amplitude","frame duration, "number of oscillations", "first amplitude" and "maximum downward speed" can distinguish between as more, and as less attentively perceived head nods. We found that head nods were generally perceived to convey more attentiveness than audio backchannels. We could furthermore show, that the fusion of modalities increases the degree of perceived attentiveness. We were able to automatically rank head-nods with an average accuracy of 74.3%. All in all it can be concluded that participants of a conversational corpus convey different levels of attentiveness through their audio-visual feedback. By using kinect

data from actual head nods, we are able to convey different degrees of attentiveness in a virtual agent. In future work we aim at using the here proposed multimodal synthesis of audio visual backchannels to better model the variation we observe in human-human conversations also in human–agent and human–robot interactions. For instance, we could focus on the behavioural synthesis of different listener categories, combining the results of this study with the results described in [14] and [13], or further exploit attention cues extracted from gaze [7] and multimodal attention [1] models to study the combination of attention and backchannels (nods, audio) on the degree of attentiveness.

8. ACKNOWLEDGMENTS

This work was partly funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF), and the H2020 European Union project MUMMER (H2020-688147). Catharine Oertel and Joakim Gustafson would like to acknowledge the support from the Horizon 2020 project BabyRobot (contract no 687831).

9. REFERENCES

[1] S. Ba and J.-M. Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008.

[2] E. Bevacqua, S. Pammi, S. J. Hyniewska, M. Schröder, and C. Pelachaud. Multimodal backchannels for embodied conversational agents. In *Intelligent Virtual Agents*, pages 194–200. Springer, 2010.

[3] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.1.13), 2009.

[4] K. Bousmalis, M. Mehu, and M. Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203–221, 2013.

[5] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 520–527. ACM, 1999.

[6] Y. Chen, Y. Yu, and J.-M. Odobez. Head nod detection from a full 3d model. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[7] K. Funes and J.-M. Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *IEEE Conf. on Computer Vision and Patter Recognition (CVPR)*, Columbus, 2014.

[8] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. Virtual rapport. In *Intelligent virtual agents*, pages 14–27. Springer, 2006.

[9] J. Liscombe, J. Venditti, and J. B. Hirschberg. Classifying subject ratings of emotional speech using acoustic features. 2003.

[10] Z. Malisz, M. Włodarczak, H. Buschmeier, S. Kopp, and P. Wagner. Prosodic Characteristics of Feedback Expressions in Distracted and Non-distracted Listeners. *The Listening Talker*, (May):36–39, 2012.

[11] Z. Malisz, M. Włodarczak, H. Buschmeier, J. Skubisz, S. Kopp, and P. Wagner. The ALICO corpus: analysing the active listener. *Journal of Language Resources and Evaluation*, in press.

[12] D. Neiberg, G. Salvi, and J. Gustafson. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication*, 55(DECEMBER):451–469, 2013.

[13] C. Oertel, J. Edlund, P. Wagner, and J. Gustafson. Gaze patterns in turn-taking. In *Proceedings of Interspeech 2012*, pages 2246–2249. ISCA, 2012.

[14] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez. Deciphering the silent participant. on the use of audio-visual cues for the classification of listener categories in group discussions. In *International Conference on Multimodal Interaction*, number EPFL-CONF-213069. ACM, 2015.

[15] C. Oertel, K. A. Funes Mora, S. Sheikhi, J.-M. Odobez, and J. Gustafson. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proc. of the workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32, 2014.

[16] C. Oertel, J. Gustafson, and A. W. Black. Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances. In *Proc. of Interspeech 2016*, pages 2915–2919, 2016.

[17] S. Pammi and M. Schröder. Annotating meaning of listener vocalizations for speech synthesis. In *3rd Int. Conf. on Affective Computing and Intelligent Interaction and Workshops*, 2009.

[18] R. Poppe, K. P. Truong, and D. Heylen. Backchannels: Quantity, type and timing matters. In *Intelligent Virtual Agents*, pages 228–239. Springer, 2011.

[19] H. M. Rosenfeld and M. Hancks. The nonverbal context of verbal listener responses. *The relationship of verbal and nonverbal communication*, 25, 1980.

[20] C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines. The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 290–296. ACM, 2006.

[21] G. Skantze and S. Al Moubayed. Iristk: A statechart-based toolkit for multi-party face-to-face interaction. In *Proc. of the 14th ACM Int. Conf. on Multimodal Interaction*, pages 69–76, 2012.

[22] G. Skantze, M. Johansson, and J. Beskow. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proc. of the Int. Conf. on Multimodal Interaction*, pages 67–74, 2015.

[23] K. P. Truong, R. Poppe, I. De Kok, and D. Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. *Proc. of Interspeech*, pages 2973–2976, 2011.

[24] P. Wagner, Z. Malisz, and S. Kopp. Gesture and Speech in Interaction: An Overview. *Speech Communication*, 57(Special Iss.), 2014.

[25] Z. Wang, J. Lee, and S. Marsella. Towards more comprehensive listening behavior: beyond the bobble head. In *International Workshop on Intelligent Virtual Agents*, pages 216–227. Springer, 2011.