User feedback in human-robot interaction: Prosody, gaze and timing

Gabriel Skantze, Catharine Oertel, Anna Hjalmarsson

KTH Speech, Music and Hearing, Stockholm, Sweden
gabriel@speech.kth.se, catha@kth.se, annah@speech.kth.se

Abstract

This paper investigates forms and functions of user feedback in a map task dialogue between a human and a robot, where the robot is the instruction-giver and the human is the instruction-follower. First, we investigate how user acknowledgements in task-oriented dialogue signal whether an activity is about to be initiated or has been completed. The parameters analysed include the users' lexical and prosodic realisation as well as gaze direction and response timing. Second, we investigate the relation between these parameters and the perception of uncertainty.

Index Terms: Feedback, Prosody, Gaze, Human-robot interaction

1. Introduction

Feedback is an essential part of human-human interaction where interlocutors continually confirm each other's utterances in order to build a common ground [1]. Improved understanding of how the prosodic realisations and timing of feedback is related to semantic and pragmatic functions in dialogue offers great potential for the development of human-like spoken dialogue systems [2]. In human-machine interaction, the understanding and generation of speech has traditionally been done in a strictly turn-based manner. However, recent efforts in the area of incremental dialogue processing [e.g. 3] have focused on how to build dialogue systems that process speech in a step-wise and parallel fashion. A dialogue system that processes speech incrementally can potentially utilize the timing as well as the production of user feedback to continuously alter its behaviour to the user's level of understanding [4,5].

Feedback is a broad term that denotes many types of verbal and non-verbal behaviours. In this paper, we focus on a subset of these behaviours, namely short acknowledgements such as “mhm”, “okay”, and “yes”. More specifically, we analyse system-directed acknowledgements in a corpus of human-robot Map Task dialogues in order understand how the users' realisations of these acknowledgements signal uncertainty as well as whether the system's route instruction has been completed. Our future motivation is to employ the discriminative features in a dialogue system to continuously classify these tokens according to functionality and determine the system's subsequent course of action.

2. Background

In human-human interaction, listeners provide short acknowledgements like “yes”, “okay” and “mhm” to continually acknowledge the speaker's incoming utterances [6]. A subset of such acknowledgements includes signals of continued attention, often referred to as continuers [7] or backchannels [8]. While carrying little propositional content and being unobtrusive in character, it has been shown that acknowledgements play a significant role in the collaborative processes of dialogue [9]. Furthermore, it has been found that the timing and frequency of acknowledgements is critical [10] and that the lexical and prosodic realisations of these tokens provide the speaker with information about the listener's attitude [11] and level of uncertainty [12,13]. Additionally, in a face-to-face setting, the speakers' gaze patterns have been shown to play an important role for the timing of backchannels [14,15].

In the area of human-machine interaction, there are several studies on how to automatically find suitable places to give acknowledgements, based on human-human [16,17,18] as well as human-machine [19] dialogue data. There are also studies on how to prosodically realise such acknowledgements [20,21]. However, systems that make use of user acknowledgements [e.g. 22,23,24,25] do not typically perform any prosodic or lexical analysis of the feedback.

In many task-oriented dialogue settings, one of the speakers has the role of an expert that guides the other speaker through some process in a step-wise manner. In this type of setting, acknowledgements do not only give the speaker information about the level of understanding, but also about whether an action has been completed. This setting is typical for many dialogue system domains, such as troubleshooting [26] and turn-by-turn navigation [27]. For such systems, it is essential to provide the instructions in appropriately sized chunks and in a timely manner, to avoid overloading the user with information and to give the user enough time to complete the requested action. In this study, we want to investigate how the user's gaze patterns as well as the prosodic realisation of acknowledgements are related to action completion in a task-oriented dialogue system.

Feedback can also reveal the speaker's level of uncertainty. Although there are several studies on how uncertainty is realised in speech in general [28,29], there are very few studies which specifically investigate acknowledgements. One exception is [12], who found that different intonation contours of cue words (e.g. "yeah", "right", "really") influence listeners' perception of uncertainty. There are also very few examples of studies of how uncertainty is expressed in human-machine dialogue. One example is [30] who adjusted the content of the output in a tutoring system based on the students' level of uncertainty. In this study, we want to investigate how the user's gaze patterns as well as the prosodic realisation of acknowledgements are related to uncertainty.

In the current study, we therefore ask the questions:

Q1: Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to discriminate between acknowledgements uttered prior to and after action completion in task-oriented human-robot interaction?

Q2: Can the users' gaze as well as their lexical and prosodic realisation of acknowledgements be used to determine the user's level of uncertainty?

3. Human-robot Map Task dialogue data

Map Task is a well establish experimental paradigm for collecting data on human-human dialogue [31]. Typically, an *instruction-giver* has a map with landmarks and a route, and is given the task of describing this route to an *instruction-follower*, who has a similar map but without the route drawn on it. The nature of this task makes it possible to collect large amounts of feedback behaviour. In a previous study [19], we have used this paradigm for collecting data on how humans elicit feedback in human-computer dialogue, where the human was the instruction-giver. In the current study, we use the same paradigm for a human-robot dialogue, where the robot is the instruction-giver and human is the instruction-follower. This has resulted in a large multi-modal corpus of user feedback behaviour.

3.1. Experimental setup

The experimental setup is shown in Figure 1. A detailed description of the setup and system is presented in [32], and we will only give a brief overview here. The user is seated opposite to the robot head Furhat [33], developed at KTH. Furhat uses a facial animation model that is back-projected on a static mask. The head is mounted on a neck, which allows the robot to direct its gaze using both eye and head movements. Between the user and Furhat is a large map printed on paper. The user can see a digital version of the map on a screen, and is given the task to draw the route Furhat describes with a digital pen on the screen. However, the landmarks on the user's screen are blurred, and the user therefore also needs to look at the large map in order to identify the landmarks. The map thereby constitutes a target for joint attention. While Furhat is describing the route, the gaze is directed at the landmarks under discussion (on the large map), which helps the user to disambiguate between landmarks. At certain places in the route descriptions, Furhat looks up at the user. The following example illustrates a typical interaction between the robot and a user, with different temporal relations between the drawing activity and the feedback:

Robot: continue towards the lights, ehm ... until you stand south of the stop lights
User: [drawing] alright [gazing at robot]
Robot: continue and pass east of the lights...
User: okay [drawing]
Robot: ...on your way towards the church
User: yes

3.2. Data collection and analysis

We collected a corpus of 24 subjects, 20 males and 4 females, interacting with the system, between the ages of 21-47. Although none of them were native speakers, all of them had a high proficiency in English. First, each subject completed a training dialogue and then three dialogues face-to-face with the robot and three dialogues when the robot's head was hidden behind a paper board, so that they couldn't see the robot but only hear its voice. The order of these conditions was systematically varied between subjects. The users' speech and face were recorded and all events in the system and the drawing activity were automatically logged. All dialogues were manually transcribed and one-word acknowledgements were identified. As acknowledgements we included all words transcribed as "okay", "yes", "yeah", "mm", "mhm", "ah", "alright" and "oh". Although there were other possible candidates for this category, their frequencies were very low. The user's gaze was also manually annotated, depending on whether the user was looking at the map, the screen or at the robot.

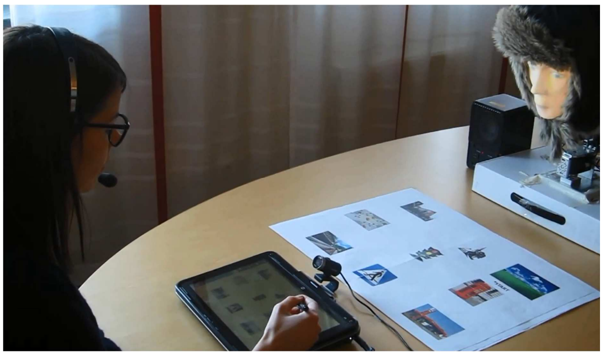

Figure 1: *The experimental setup.*

For each acknowledgement, we extracted the pitch using ESPS in Wavesurfer/Snack [34] and converted the values into semitones. In order to get a measure of **pitch slope**, we calculated the difference between the average of the second half of these values and the average of the first half for each token (i.e. negative=falling, positive=rising). Each value was then z-normalised based on the overall data for the speaker. Using these values, the **average (normalised) pitch** was calculated for each token. A measure of **duration** was also calculated by counting the number of voiced frames (each 10ms) for the token. To measure the **average (normalised) intensity**, we used Praat [35] and then calculated the average dB (z-normalised for the speaker). The intensity measures used were extracted from voiced intervals only.

For each token, we also defined a binary feature of whether the **user gazed at the robot** at any point in a time window between 1 second before and 1 second after the token. To get a measure of **drawing activity** in relation to the feedback, we calculated the number of pixels drawn between the end of the last system instruction up to the middle of the feedback, and the number of pixels drawn in a window between the middle of the feedback and 3 seconds after.

4. Results

In total, there were 1568 feedback tokens in the whole dataset. The prosodic extraction failed for some tokens, so for the prosodic analysis we were able to use 1464 tokens. For the gaze analysis, we could only use the dialogues with a face-to-face condition, which resulted in 807 tokens. In the following analysis, all tests for significance are done using two-tailed tests at the .01 level.

4.1. Task activity completion

First, we investigated how the realisation of acknowledgements relates to the user's drawing activity. Here we try to discriminate between acknowledgements with four different functions: (1) that an activity is about to be initiated (**before drawing**), (2) that it is underway (**while drawing**), (3) that it has been completed (**after drawing**), or (4) that it has already been completed in a previous step (**no drawing)**. All pen movements of more than 50 pixels within the before/after windows (as defined above) were considered as drawing. Figure 2 illustrates the timing of acknowledgements in relation to the end of the previous system utterance. A bit surprisingly, the timing of the acknowledgement does not in itself reveal very much about its temporal relation to the drawing activity, especially for early acknowledgements (which is also most common).

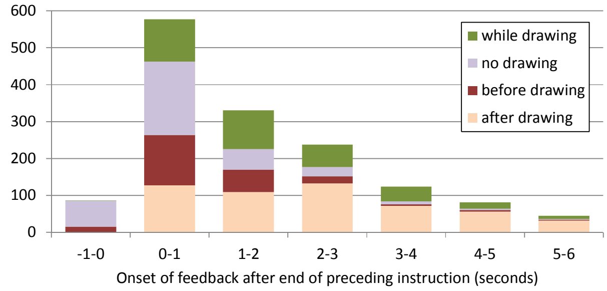

Figure 2: *Acknowledgement frequency and activity completion over response time.*

A chi-square test was used to investigate the relationship between the lexical realisation of acknowledgements and drawing activity. The results were significant ($\chi^2(21, N=1568) = 248.19$), showing that there is a difference in the distribution of lexical tokens between the different drawing activity classes. For example, as illustrated in Figure 3, "yes" is more likely to signal "no drawing" activity, whereas "okay" shows an opposite distribution.

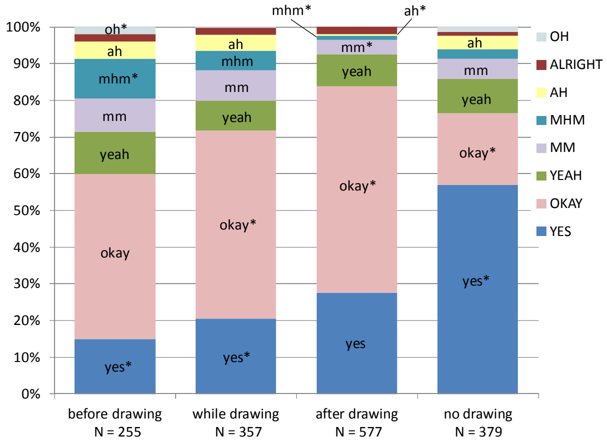

Figure 3: *Distribution of lexical tokens depending on activity completion. * marks significant differences from the overall distribution.*

Next, we used a MANOVA to explore the relationship between the prosodic realisation of the acknowledgement and the different drawing activity classes. The prosodic features described in 3.2 were used as dependent variables and the drawing activity class was used as the independent variable. The results show a general significant effect ($F(28, 5240)=39.96$; Wilk's $\Lambda=0.50$), as well as several individual significant effects in post-hoc tests (Tukey HSD). All prosodic features except average pitch showed effects. Significant results are summarized in Table 1 and Table 2. For example, acknowledgements with "no drawing" have a higher intensity, shorter duration and a relatively flat pitch. Acknowledgements "before drawing" and "while drawing" both have a lower intensity and longer duration, but differ in that "while drawing" has a higher rising pitch. Since the uneven distribution of lexical tokens may influence these differences, we also conducted separate MANOVA analyses for the two most frequent tokens – "okay" and "yes" – which also showed significant effects. As can be seen in the post-hoc tests reported in Table 2, the general patterns are the same, even though some individual effects are different.

Table 1. *The relationship between drawing activity and prosody.*

	Before Drawing	**While Drawing**	**After Drawing**	**No Drawing**
N	238	334	549	343

Intensity (z-score)	M = -0.19
	SD = 0.64
		M = -0.08
		SD = 0.53
			M = 0.00
			SD = 0.52
				M = 0.25
				SD = 0.69

Duration (v. frames)	M = 27.30
	SD = 12.97
		M = 28.11
		SD = 12.51
			M = 22.19
			SD = 9.68
				M = 18.97
				SD = 10.91

Pitch Slope (semitones)	M = 0.97
	SD = 3.05
		M = 1.98
		SD = 3.14
			M = 1.76
			SD = 3.21
				M = 0.35
				SD = 3.07

Table 2. *Post-hoc analysis for differences in effect of drawing activity on prosody.*

Token	**Parameter**	**Post-hoc effects**
All	Intensity	No > (After, While, Before) After > Before
	Duration	(While, Before) > After > No
	Pitch Slope	(While, After) > (Before, No)
"yes"	Intensity	No > (While, After)
	Duration	While > (Before, After, No)
	Pitch Slope	While > No
"okay"	Intensity	*No significant differences*
	Duration	While > (After, No)
	Pitch Slope	After > (Before, No)

For the face-to-face conditions, we also counted the number of times that the user gazed at the robot in vicinity of the feedback and a chi-square test was used to investigate the relationship between gaze and drawing activity. The results are presented in Table 3, showing clear significant effects ($\chi^2(3, N=807) = 55.99$) – users tend to look more at the robot when they do not have to draw, and even more when they have completed the drawing activity.

Table 3. *The relationship between drawing activity and gaze.*

	Before Drawing	**While Drawing**	**After Drawing**	**No Drawing**
N	131	151	321	204
Gaze at robot	34.4% SR = -2.9	37.1% SR = -2.7	66.0% SR = 3.3	55.4% SR = 0.5
	Post-hoc: After > No > Before, While			

4.2. Uncertainty

In order to analyse uncertainty in acknowledgements, a binary distinction between *certain* and *uncertain* was made, and all acknowledgment tokens were manually annotated for these two categories, resulting in 1376 certain tokens and 192 uncertain tokens. In order to validate the annotation scheme, we randomly selected 110 pairs of certain/uncertain tokens in the same lexical category. These were then presented to a second and third annotator who had to select which one in each pair sounded more certain. The annotators did not get any dialogue context and were not given information about the other annotators' labelling. The cross-annotator agreement was measured in a multirater kappa analysis [36], which resulted in a kappa score of 0.63 – a substantial agreement according to [37].

A chi-square test was used to investigate the relationship between the lexical realisations and uncertainty. The results were significant ($\chi^2(7, N=1568) = 272.87$), showing that there is a difference in the distribution of lexical tokens between the certain and uncertain categories. As illustrated in Figure 4, the distribution varies a lot – all lexical tokens except “alright” and “okay” are significantly different between the *certain* and *uncertain* categories.

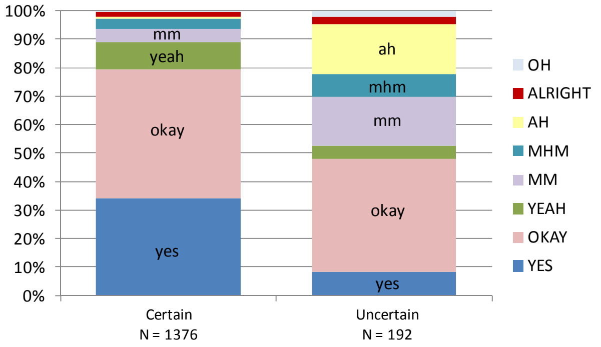

Figure 4: *Distribution of lexical tokens depending on uncertainty.*

Next, we wanted to see how uncertainty is related to the prosodic features described in 3.2, as well as the drawing activity. All these features were used as dependent variables in a MANOVA, which showed a general significant effect (F(6, 1457)=71.85; Wilk's Λ=0.77). The results are summarized in Table 4. All prosodic features except average pitch show effects. “Certain” acknowledgements have higher intensity, shorter duration, and a more rising pitch. Table 4 also shows a separate analysis of “okay” (which was equally common in both categories and frequent enough). As can be seen, all significant differences except pitch slope remain. However, for intensity the trend now goes in the opposite direction – “okay” has a lower intensity when expressed as “certain”. We do not have a good explanation for this, and this difference needs to be investigated further.

As can be seen in Table 4, the drawing-activity before users produce an acknowledgement is higher when they are perceived as "certain", whereas their drawing activity after they produced the feedback token is lower when they are perceived as "certain". One possible explanation for this could be that when users do not know immediately what to draw, they are more likely to express this uncertainty before drawing.

Table 4. *Significant differences in prosody and drawing activity between certain/uncertain.*

	All tokens		**“okay”**	
	Certain	**Uncertain**	**Certain**	**Uncertain**
N	1280	184	595	70

Intensity
(z-score)
	M=0.08
	SD=0.57
		M=-0.47
		SD=0.64
			M=-0.83
			SD=0.45
				M=-0.53
				SD=0.55

Duration
(v. frames)
	M=22.01
	SD=9.80
		M=34.71
		SD=17.16
			M=23.44
			SD=6.78
				M=27.96
				SD=12.68

Pitch Slope
(semitones)
	M=1.43
	SD=3.24
		M=0.82
		SD=2.90
			Not significant	

Draw
before
	M=162.4
	SD=191.0
		M=99.6
		SD=117.3
			M=202.3
			SD=194.7
				M=115.9
				SD=126.4

Draw
after
	M=69.5
	SD=191.0
		M=157.4
		SD=168.8
			M=77.7
			SD=129.3
				M=166.7
				SD=173.1

For the face-to-face conditions, we also counted how many times the user gazed at the robot in vicinity of the acknowledgement, depending on uncertainty. The results are shown in Table 5, and show significant effects (χ^2(1, N=807) = 23.18) – users tend to look more at the robot when they are certain than when they are uncertain.

Table 5. *Certain/uncertain vs. user’s gaze.*

	Certain	**Uncertain**
N	710	97
Gaze at robot	55.9% SR = 1.1	29.9% SR = -3.1

5. Conclusions and Discussion

In this paper we investigated forms and functions of user acknowledgements in a map task dialogue between a human and a robot. First, we noticed that users’ realisations of acknowledgements in task-oriented dialogue can be used to identify whether an activity is about to be initiated, is underway, has been completed, or was already completed in a previous step. An analysis of response time showed that this factor in itself does not reveal much about task completion. However, analyses of the distribution of lexical tokens, prosody and gaze, showed that these factors reveal a lot about task activity completion. For example, the use of "okay" is associated with execution of the action, whereas "yes" is more likely to signal that the activity has already been completed. The latter can also be signalled with a higher intensity. The use of "mhm" is more common before or while executing an action, but this can also be signalled with a longer duration in other lexical tokens. In addition, users tend to gaze more at the robot when the action has been completed, which is in line with general gaze patterns found in turn-taking behaviour [38,15].

Next, we investigated the relation between these parameters and the perception of uncertainty. The results show that uncertainty is more often expressed with “mm” and “ah”, whereas certainty is more often expressed with “yes”. In the cases where the distributions are similar, as for “okay”, the prosodic patterns could be used to identify uncertain user feedback. For example, uncertainty seems to be associated with longer duration, a finding which is line the analysis reported in [11]. It was also shown that the users’ gaze and drawing activity (if observable by the system) is informative. The significant effect of drawing activity also indicates an interesting relationship between uncertainty and task completion that calls for further investigation.

As a next step we will use these features in combination to build a classifier for task activity completion and uncertainty that can be used online in a dialogue system. For example, the speech generation component could incrementally adapt the output to the user’s feedback [4,5]. The problem of knowing whether an action has been completed is of course not limited to drawing a route on a map, but should be applicable to many types of task-oriented dialogue settings.

To our knowledge, no other studies have previously explored how users’ gaze patterns and prosodic realisations of acknowledgements are related to task completion and uncertainty in a human-machine dialogue setting.

6. Acknowledgements

Gabriel Skantze is supported by the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (#2011-6237). Catharine Oertel is supported by *GetHomeSafe* (EU 7th Framework STREP 288667). Anna Hjalmarsson is supported by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation*.

7. References

[1] Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

[2] Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

[3] Schlangen, D., & Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse, 2*(1), 83-111.

[4] Skantze, G., & Hjalmarsson, A. (2012). Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language, 27*(1), 243-262.

[5] Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., & Schlangen, D. (2012). Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of SigDial* (pp. 295-303). Seoul, South Korea.

[6] Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259-294.

[7] Schegloff, E. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In Tannen, D. (Ed.), *Analyzing Discourse: Text and Talk* (pp. 71-93). Washington, D.C., USA: Georgetown University Press.

[8] Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.

[9] Schober, M., & Clark, H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

[10] Poppe, R., Truong, K., & Heylen, D. (2011). Backchannels: Quantity, Type and Timing Matters. In *11th International Conference, IVA 2011* (pp. 228-239). Reykjavik, Iceland.

[11] Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Proceedings of Speech Prosody* (pp. 325-328).

[12] Lai, C. (2010). What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. In *Proceedings of Interspeech*. Makuhari, Japan.

[13] Neiberg, D., & Gustafson, J. (2012). Cues to perceived functions of acted and spontaneous feedback expressions. In *The Interdisciplinary Workshop on Feedback Behaviors in Dialog*.

[14] Bavelas, J., Coates, L., & Johnson, T. (2002). Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication, 52*(3), 566-580.

[15] Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., & Gustafson, J. (2012). Gaze Patterns in Turn-Taking. In *Proc. of Interspeech 2012*. Portland, Oregon, US.

[16] Ward, N., & Tsukahara, W. (2003). A study in responsiveness in spoken dialog. *International Journal of Human-Computer Studies, 59*, 603-630.

[17] Poppe, R., Truong, K., Reidsma, D., & Heylen, D. (2010). Backchannel Strategies for Artificial Listeners. In *10th International Conference, IVA 2011* (pp. 146-158). Philadelphia, Pennsylvania, USA.

[18] Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest.

[19] Skantze, G. (2012). A Testbed for Examining the Timing of Feedback using a Map Task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Portland, OR.

[20] Wallers, Å., Edlund, J., & Skantze, G. (2006). The effects of prosodic features on the interpretation of synthesised backchannels. In André, E., Dybkjaer, L., Minker, W., Neumann, H., & Weber, M. (Eds.), *Proceedings of Perception and Interactive Technologies* (pp. 183-187). Springer.

[21] Stocksmeier, T., Kopp, S., & Gibbon, D. (2007). Synthesis of prosodic attitudinal variants in german backchannel ja. In *Proceedings of Interspeech 2007*.

[22] Iwase, T., & Ward, N. (1998). Pacing spoken directions to suit the listener. In *Proceedings of ICSLP* (pp. 1203-1207). Sydney, Australia.

[23] Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)* (pp. 553-561).

[24] Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.

[25] Buschmeier, H., & Kopp, S. (2011). Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback. In *Proceedings of IVA* (pp. 169-182). Reykjavik, Iceland.

[26] Boye, J. (2007). Dialogue management for automatic troubleshooting and other problem-solving applications. In *Proceedings of the 8th SIGDial workshop on discourse and dialogue*. Antwerp, Belgium.

[27] Boye, J., Fredriksson, M., Götze, J., Gustafson, J., & Königsmann, J. (2012). Walk this way: Spatial grounding for city exploration. In *IWSDS2012 (International Workshop on Spoken Dialog Systems)*.

[28] Liscombe, J., Venditti, J., & Hirschberg, J. (2006). Detecting Question-Bearing Turns in Spoken Tutorial Dialogues. In *Proceedings of Interspeech 2006, Pittsburgh, PA, USA*.

[29] Pon-Barry, H. (2008). Prosodic Manifestations of Confidence and Uncertainty in Spoken Language. In *Proceedings of Interspeech* (pp. 74–77). Brisbane, Australia.

[30] Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication, 53*(9-10), 1115–1136.

[31] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

[32] Skantze, G., Hjalmarsson, A., & Oertel, C. (submitted). Exploring the effects of gaze and pauses in situated human-robot interaction. Submitted to *SigDial*.

[33] Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multi-party Interaction. *International Journal of Humanoid Robotics, 10*(1).

[34] Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

[35] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

[36] Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. In *Joensuu University Learning and Instruction Symposium*. Joensuu, Finland.

[37] Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

[38] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, 22-63.