Making Turn-Taking Decisions for an Active Listening Robot for Memory Training

Martin Johansson[1(✉)], Tatsuro Hori[2], Gabriel Skantze[1], Anja Höthker[3], and Joakim Gustafson[1]

[1] KTH Speech, Music and Hearing, Stockholm, Sweden
vhmj@kth.se, {gabriel,jocke}@speech.kth.se
[2] Toyota Motor Corporation, Toyota, Japan
tatsuro_hori@mail.toyota.co.jp
[3] Toyota Motor Europe, Brussels, Belgium
Anja.Hoethker@toyota-europe.com

Abstract. In this paper we present a dialogue system and response model that allows a robot to act as an active listener, encouraging users to tell the robot about their travel memories. The response model makes a combined decision about when to respond and what type of response to give, in order to elicit more elaborate descriptions from the user and avoid non-sequitur responses. The model was trained on human-robot dialogue data collected in a Wizard-of-Oz setting, and evaluated in a fully autonomous version of the same dialogue system. Compared to a baseline system, users perceived the dialogue system with the trained model to be a significantly better listener. The trained model also resulted in dialogues with significantly fewer mistakes, a larger proportion of user speech and fewer interruptions.

Keywords: Turn-taking · Active listening · Social robotics · Memory training

1 Introduction

Social robots of the future are envisioned to assist us in our daily life in various ways. One interesting field where social robots could be of use concerns elderly care and improving the life of elderly people. In this field, there is for example work towards personalized conversational interfaces [1] and embodied conversational agents (ECA) that can assist elderly in organizing their daily activities [2], and serve as conversational companions for people with dementia [3, 4]. Aside from the possibility of providing a conversational partner that does not mind repetitions, such a system could potentially also be used for memory training and monitoring. A key challenge in realizing such a system is to create an artificial agent that can be characterized as a good listener. The aim of this paper is to develop a conversational system for a social robot that acts as an active listener to a user talking about memories of past events. Using machine learning, we train a model that can predict *when* the system should respond, and *how* it should respond, based on Wizard-of-Oz data.

A. Agah et al. (Eds.): ICSR 2016, LNAI 9979, pp. 940–949, 2016.
DOI: 10.1007/978-3-319-47437-3_92

2 Background

In this work we are interested in a dialogue scenario where the main flow of information goes from a speaker to a single listener. Despite the imbalance in flow of information between the interlocutors, the behavior of the listener is still important for the speaker. Experiments on active listening have shown that when speakers receive more feedback, their narratives become more comprehensible, and that listener feedback helps to coordinate what the speaker says with what the listener needs to know [5]. A common form of feedback in such settings is *backchannels* [6] – short acknowledgement utterances, such as “uh-huh” or “yeah”, or non-verbal gestures, such as head nods. By producing a backchannel, the listener does not claim the floor, but rather encourages the speaker to continue. However, a good listener needs to balance the use of such backchannels with other types of responses, such as follow-up questions [7].

Human speakers in dialogue synchronize their turn-taking to minimize gaps and overlaps [8] through the use of turn-holding and turn-yielding cues, including prosody, syntax and gestures [9–11], as well as gaze in face-to-face interactions [12]. The more turn-yielding cues a speaker presents together, the higher the likelihood for the listener to take the turn. For example, flat final pitch, syntactic incompleteness and filled pauses are strong cues to turn hold. Meena et al. [13] presented a data-driven model that used automatically extracted features from syntax, prosody and context to detect suitable response locations in a spoken dialogue system that was listening to a user giving route descriptions. It was found that syntax was the most important feature, given that the speech recognition (ASR) worked well. When ASR performance dropped, prosody and context were also viable features. For our setting, this is an encouraging finding, since speech recognition of spontaneous travel memories is potentially very challenging, and we want the system to be a good listener even without using an ASR. In this paper, we extend the work by Meena et al. by not only detecting suitable response locations; we also want the system to be able to choose the type of response to make: a backchannel acknowledgment or taking the turn and asking a follow-up question.

The use of data-driven methods for creating a dialogue control component for an active listener has also been investigated by Meguro et al. [14], who built a model based on sequences of manually annotated dialogue acts (DA) in human-human textual dialogues. However, their model was not evaluated in a real spoken dialogue setting. In the field of Embodied Conversational Agents (ECAs), there have also been works on modelling active listening behavior [15, 16]. These studies have found that people interact with a responsive listener longer than with an unresponsive one, and also use more words. Sakai et al. [3] proposed an ECA as a conversational partner for individuals with dementia, using a combination of questions asked by the ECA and a rule-based system to generate verbal and non-verbal backchannel feedback, based on silence length and pitch. Yasuda et al. [4] also developed and evaluated an ECA for this use, and found the ECA to elicit utterances from subjects with Alzheimer corresponding to 74 % of the length of those in an identical human-human condition with the same subjects. However, none of these studies were done in a human-robot interaction scenario. Moreover, they did not propose a complete data-driven model of *when* to respond and *what type* of response to give (such as a backchannel or a follow-up question).

3 A Listening Social Robot

The domain chosen for the human-robot dialogue in this paper is a travel memory domain, where subjects are to tell the robot about a past visit to a foreign country, while the robot listens actively to elicit more elaborate descriptions. The interaction is dyadic with the subject seated in front of the robot, as illustrated in Fig. 1.

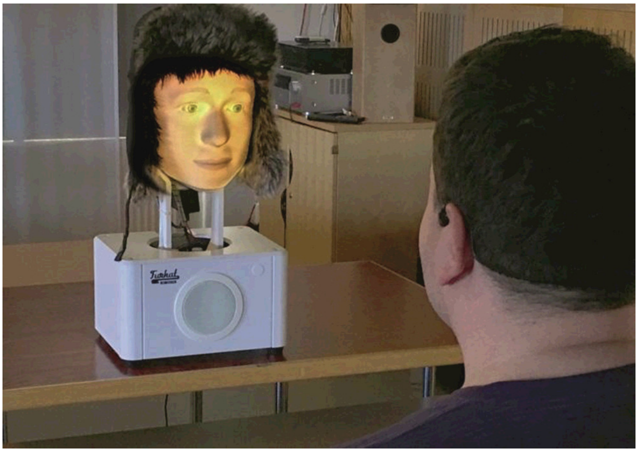

Fig. 1. Illustration of the experiment setting with the robot Furhat and a user interacting

Furhat [17], shown in Fig. 1, was chosen as the robot participant in the experiment. Furhat is a back-projected human-like robot head using speech synthesis and state-of-the-art facial animation, mounted on a mechanic neck. The facial animation architecture allows for speech with accurate synchronized lip movements, as well as eye movement and facial expressions.

3.1 The Listening Dialogue System

The interaction between a subject and the system comprise three phases, starting with an introductory phase where the system holds the initiative to get the conversation started. The phase begins with the system greeting the subject and asking for the subject's name, which country the travel memories will concern, and ends with the system asking which city the subject visited. The introductory phase is seamlessly followed by the active listening phase, where the subject is asked to tell the system about travel memories from the selected country. The flow of the active listening phase is illustrated in Fig. 2. As the goal of the system is to keep the subject talking about recalled travel memories rather than to extract specific information, the system is designed to allow the subject to keep the initiative and to keep the flow of information going mainly from the subject to the system, using backchannels and questions to encourage the user to continue speaking.

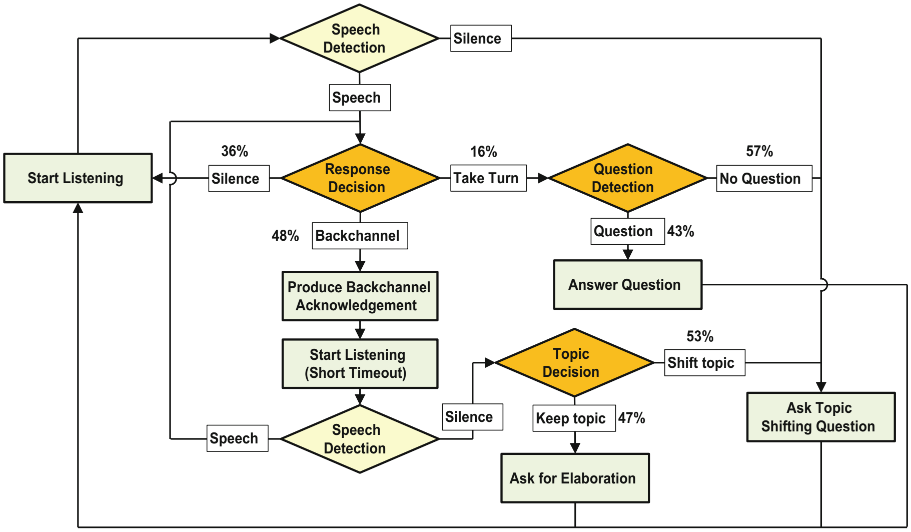

Fig. 2. The active listening phase of the listening dialogue system. Percentages indicate decision distributions in the training corpus.

The active listening flow is centered on the voice activity detection (VAD) component of the system, considering the end of detected voice activity as a potential location for making a response. A **Response decision** module then makes the decision to either stay silent, to produce a verbal backchannel acknowledgment, or to make a longer response. Since the user sometimes might ask counter-questions, a **Question detection** module is used to either react to the question or to ask a follow-up question that will advance the dialog to a new (sub) topic. If no speech is detected, the latter action will automatically be selected; the system will ask the subject a question related to the country under discussion. In this first version of the system, the system picked the utterance from a pre-defined handcrafted list, and the choices of countries to talk about were limited to three: France, Germany and Spain. If no speech is detected shortly after a produced verbal backchannel acknowledgement, a **Topic decision** module makes the decision to stay on topic by asking for an elaboration or to advance to a new topic as described above.

Finally, when a predefined amount of time has passed, the active listening phase ends and is replaced by the closing phase where the system retakes the initiative, ends the conversation and thanks the subject for the chat.

The system was implemented using the open source dialogue system framework IrisTK [18], which is used to control the dialogue as well as the non-verbal aspects of the robot. The gaze behavior was designed to follow general rules found in studies on human-human interaction [12]: When the robot was about to speak a longer utterance, it gazed away from the user at the beginning of the utterance and gazed back towards the end. In all other situations, the robot gazed towards the user, with small saccadic movements.

3.2 Interaction with the Robot

Before an interaction between a subject and the robot started, the task was explained by the experiment leader. The subject was then seated in front of the robot, illustrated in Fig. 1, and equipped with a headset used for audio recording and the system's speech recognition. An example from an interaction is presented in Table 1, containing both cases where a backchannel acknowledgment from the system elicited more information from the user (such as R.15), and cases where no additional information was gained (such as R.6), and where a follow-up question would probably have been more appropriate. There are also examples of the robot remaining silent during a filled pause from the user (S.4–S.5), reacting to a question from the user (R.21), as well as staying on topic by asking for an elaboration (R.11) and advancing to a new topic (such as R.8).

Table 1. Excerpt from a dialogue between the robot (R) and a subject (S).

Speaker	Utterance	Response type
R.1	OK. Let's talk about Germany. Which city did you visit?	RequestCity
S.2	I visited Berlin.	
R.3	Could you tell me about that?	Elaboration Question
S.4	It was really nice and um …	
S.5	I walked a lot.	
R.6	ah	Backchannel
S.7	yeah	
R.8	ehm … Do you think I should go there?	Topic-shift Question
S.9	Yeah, sure.	
R.10	oh	Backchannel
R.11	Why do you think so?	Elaboration Question
S.12	It is really nice. I think you definitely should go there.	
R.13	I heard you can eat nice sausage there. Did you try that?	Topic-shift Question
S.14	In Berlin, yes. It was really nice eating sausage in Berlin.	
R.15	mhm	Backchannel
S.16	Sauerkraut	
R.17	oh	Backchannel
S.18	yeah	
R.19	ehm … Did you see more interesting things there?	Topic-shift Question
S.20	Yeah, they had a lot of graffiti. Do you like graffiti?	
R.21	ehm … I will answer next time we meet.	Question Reaction

3.3 Wizard of Oz Data Collection and Annotation

A round of initial data collection was carried out through a Wizard-of-Oz version of the system, where the wizard was deciding the verbal actions of the robot during the active listening phase illustrated in Fig. 2, thus playing the role of the three decision models we wanted to train. To reduce the latency of the robot's turn-taking, the wizard claimed the floor through a hesitation sound (e.g., "uhm") while choosing the type of response.

Interactions with five subjects were used to tune the system as well as to train the wizard, while another five subjects participated in the initial data collection, recording one dialogue each.

The recorded audio was automatically segmented into Inter Pausal Units (IPUs), using energy-based VAD with a maximum of 500 ms internal silence. For the annotation, we considered the end of IPUs as potential response locations, and the annotator decided whether the best reaction would be to either stay **silent**, produce a **backchannel** or **take the turn**. The first two dialogues were annotated by two of the authors with a good inter-annotator agreement, a kappa score of 0.72. The remaining three dialogues were then annotated by one author. In total, a set of 131 decision points from the 5 dialogues were annotated for turn-taking decisions. The first decision, to stay *silent* (36 % of annotated instances) represents instances where the robot should not say anything. The second decision, *backchannel* (48 %), represents instances where the robot should make a backchannel (ex. uh-huh) to promote more elaborate descriptions. The final decision, *take the turn* (16 %), represents for example instances where the robot should ask a question to encourage a slight topic switch, since the user did not appear to intend continue speaking on the current topic (ex. “What did you like best?” or “Interesting, why do you say that?”) or react to a question from the user.

3.4 A Data-Driven Model for Response Decisions

The human-robot travel memories corpus collected and annotated in Sect. 3.3 was used to train a model for making response decisions in the dialogue. The model was based on the Random Forest Algorithm in the WEKA toolkit [19], using only automatically extractable features as outlined below. In total, 15 features were used to represent context and prosody. For context, we used the length of the user’s turn so far and three features from the dialogue manager: the previous response decision (silent, backchannel or sentence), the amount of time since the system last had the turn and finally what type of sentence the system uttered during that turn (answering a question from the user, asking an open question, or asking the user to elaborate). As prosodic features, we used final pitch and energy. A pitch tracker based on the Yin algorithm [20] was used to estimate the F_0 at a rate of 100 frames per second. The F_0 values were then transformed to log scale and z-normalized for each user. For each IPU, the last voiced frame was identified and then regions of **200 ms** and **500 ms** ending in this frame were selected. Based on this, we used 3 energy features (the maximum energy for the 200 ms and 500 ms regions and for the full IPU) and 8 pitch features (mean, standard deviation and slope of the normalized F_0 values in regions of both the last 200 ms and 500 ms, maximum of last 500 ms, the maximum and standard deviation of the normalized F_0 values over the full IPU).

The weighted F-score of this model using 10-fold cross-validation was 0.65, compared to 0.31 for a model always selecting the majority class. Prosody was more useful than context; a model using pitch and energy yielded a weighted F-score of 0.62 compared to 0.52 for a model using only dialogue context.

The corpus was also used to build a Question detection model and a Topic decision model for use as illustrated in Fig. 2, using the JRIP Algorithm in the WEKA toolkit.

For the Question detection model, a non-falling final pitch was selected as an indicator for questions. For the Topic decision model, if the user utterance ended in a low pitch, and the system had recently made a topic shift question, the system should not shift topic again. These models performed better than the majority class baselines, but since there are only a small amount of examples for these models in the corpus, 21 and 34 respectively, we will focus our analysis primarily on the Response decision model.

4 Results and Discussion

The dialogue system described in Sect. 3 was evaluated using 15 subjects (14 male users and 1 female user, 25–62 years old) from among employees and students at KTH, who each interacted with both a baseline system and the trained system. It is not obvious what to use as a baseline system. We chose to use a random decision model, where the options where weighted based on the distributions of the decisions in the annotated training material. The order of interaction with the two systems was shuffled to avoid ordering effects, resulting in 8 of the subjects experiencing the baseline system first and 7 subjects the new system first. Each interaction was 4 min long, controlled by the system.

4.1 Qualitative Evaluation

The developed system and the baseline system were evaluated qualitatively by letting the subjects fill out a questionnaire, using a semantic differential scale. After each interaction, subjects were asked to rate five aspects of the interaction: (i) how exciting it was, (ii) if the turn-taking was bad or good, (iii) the quality of the content in the robot's feedbacks, (iv) the naturalness of the robot's gaze behavior, and (v) if the system was perceived as a good listener. The mean rating for each aspect and system is shown in Fig. 3 to illustrate subjects' ratings of the systems. As can be seen, the subjects' ratings of the trained system as a good listener was statistically significantly higher than that of the baseline system (Wilcoxon Signed Ranks Test, $Z = -2.783$, $p = .0054$). The clear difference between the two systems is interesting considering the closer ratings of the other aspects. One possible explanation to this is that being a good listener is an amalgam of different skills.

It is also apparent that there is room for improvement of all aspects. However, it is not clear how to interpret the absolute ratings for a specific aspect, as there were for example no comparisons with human listeners. Thus, we only compare the ratings of the trained system with those of the baseline system, although we note that the automated gaze behavior (same for both systems) received high ratings.

4.2 Perception Test

To complement the participating subjects' ratings, we carried out a perception test where three subjects, who did not take part in the interactions, listened to recorded audio from the interactions with the task of spotting bad interactions. The subjects were instructed

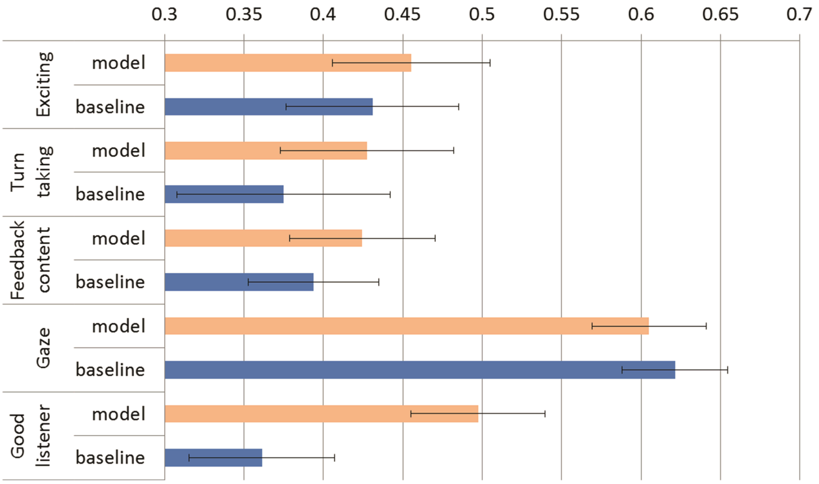

Fig. 3. Mean of questionnaire results encoded on a scale from 0 (bad) to 1 (good), with error bars illustrating the estimated standard error of the mean.

to push a button whenever the system made an inappropriate decision, either by remaining silent when a response was expected, by making a response at the wrong place, or by making the wrong type of response. The trained system was marked for significantly fewer inappropriate decisions than the baseline system (*M* = 7.7, *SD* = 4.1 vs. *M* = 10.3, *SD* = 2.6) (one-sided T-test, p = .024).

In addition to the mistakes of the Response decision model, the annotators marked 23 decisions for the baseline system and 19 decisions for the trained system as problematic, despite the systems' choices of a correct type of action. These represent errors made by the Question detection and Topic decision models. This could for example be due to the system not recognizing that the subject had asked it a question, or asking a question that did not fit the context. These kinds of errors are arguably detrimental to the illusion of the system as a good listener who understands the speaker, and indicates that this kind of shallow processing needs to be complemented with some sort of language understanding.

4.3 Quantitative Evaluation

We also carried out a quantitative evaluation of the collected dialogue data, namely the proportion of subject speech and the number of interruptions of the user made by the system. The evaluation was based on automated segmentation of the recorded speech using energy-based VAD. The proportion of subject speech time of the total speech time in each dialogue was significantly higher for the trained system (*M* = 54.6 %, *SD* = 10.7 %) than for the baseline system (*M* = 49.0 %, *SD* = 6.2 %) (one-sided T-test, p = .0455), while the number of times the system interrupted the subject in each dialogue were lower (*M* = 6.0, *SD* = 2.9 vs. *M* = 8.3, *SD* = 4.4) (one-sided T-test, p = .0496).

5 Conclusions and Future Work

In this paper, we have presented a dialogue system and response model that allows a robot to act as an active listener, encouraging users to tell the robot about their travel memories. Unlike previous active listening models, the system makes a combined decision about when to respond and what type of response to give, thereby eliciting more elaborate descriptions from the user. The model was trained with human-robot interactions collected through a Wizard-of-Oz setting, and despite the relatively small training data set of five interactions, the trained system was perceived by subjects as a better listener than a weighted random baseline system, and the trained system also made fewer problematic decisions.

There are a number of possible improvements that can be made. The current system for example only gives verbal feedback to the speaker when listening, and tries to avoid feedback within utterances. Based on the results of Gratch et al. [15], adding non-verbal feedbacks, also within utterances, could be beneficial, and while the current system use gaze to signal turn-holding as described by [12], it does not make use of the speaker's gaze. The current system also makes mistakes due to lack of understanding, for example by choosing irrelevant prompts or missing that the user asked the system a question.

Given the small amount of data used for training the models, and the fact the we only used context and prosody for the models, the results are encouraging. For future work, we will add speech recognition to the models, not only to improve the turn-taking behavior, but also to extract the contents of the travel memories. We think that the system could be valuable for supporting and diagnosing people with dementia, and this is something we also want to investigate in the future.

Acknowledgements. This research was funded by Toyota Motor Corporation, as well as the Swedish research council (VR) project *Coordination of Attention and Turn-taking in Situated Interaction* (2013-1403).

References

1. Benyon, D., Mival, O.: Introducing the companions project: intelligent, persistent, personalised interfaces to the internet. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI…But Not As We Know It, vol. 2, pp. 193–194 (2007)
2. Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., Tobiasson, H.: The MonAMI reminder: a spoken dialogue system for face-to-face interaction. In: Interspeech 2009, Brighton, U.K. (2009)
3. Sakai, Y., Nonaka, Y., Yasuda, K., Nakano, Y.I.: Listener agent for elderly people with dementia. In: HRI 2012, pp. 199–200 (2012)
4. Yasuda, K., Aoe, J., Fuketa, M.: Development of an agent system for conversing with individuals with dementia. In: The 27th Annual Conference of the Japanese Society for Artificial Intelligence (2013)
5. Kraut, R.E., Lewis, S.H., Swezey, L.W.: Listener responsiveness and the coordination of conversation. J. Pers. Soc. Psychol. **43**(4), 718–731 (1982)

6. Yngve, V.H.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, Chicago, pp. 567–578 (1970)
7. Kobayashi, Y., Yamamoto, D., Koga, T., Yokoyama, S., Doi, M.: Design targeting voice interface robot capable of active listening. In: 5th ACM/IEEE International Conference on Human-robot Interaction, pp. 161–162 (2010)
8. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50**, 696–735 (1974)
9. Duncan, S.: Some signals and rules for taking speaking turns in conversations. J. Pers. Soc. Psychol. **23**(2), 283–292 (1972)
10. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. Lang. Speech **41**, 295–321 (1998)
11. Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. **25**(3), 601–634 (2011)
12. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychol. **26**, 22–63 (1967)
13. Meena, R., Skantze, G., Gustafson, J.: A data-driven model for timing feedback in a map task dialogue system. In: 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France, pp. 375–383 (2013)
14. Meguro, T., Higashinaka, R., Minami, Y., Dohsaka, K.: Controlling listening-oriented dialogue using partially observable markov decision processes. In: Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, PA, USA, pp. 761–769 (2010)
15. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S.C., Morales, M., van der Werf, R.J., Morency, L.-P.: Virtual rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
16. Huang, L., Morency, L.-P., Gratch, J.: Virtual rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
17. Al Moubayed, S., Skantze, G., Beskow, J.: The furhat back-projected humanoid head - lip reading, gaze and multiparty interaction. Int. J. Humanoid Rob. **10**(1), 1350005 (2013)
18. Skantze, G., Al Moubayed, S.: IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In: Proceedings of ICMI, Santa Monica, CA (2012)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
20. de Cheveigné, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002)