Estimating Speaker's Engagement from Non-verbal Features Based on an Active Listening Corpus

Lei Zhang[1], Hung-Hsuan Huang[1,2(✉)], and Kazuhiro Kuwabara[1]

[1] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan
hhhuang@acm.org
[2] Center for Advanced Intelligence Project, RIKEN, Kyoto, Japan

Abstract. The elderly who live alone are increasing rapidly in these years. For their mental health, maintaining their social life with others is reported useful. Our project aims to develop a listener agent who can engage active listening dialog with the elderly users. Active listening is a communication technique that the listener listens to the speaker carefully and attentively. The listener also ask questions for confirming or showing his/her concern about what the speaker said. For this task, it is essential for the agent to evaluate the user's engagement level (or the attitude) in the conversation. In this paper, we explored an automatic estimation method based on empirical results. An active listening conversation experiment with human-human participants was conducted for corpus collection. The speakers' engagement attitude in the corpus was subjectively evaluated by human evaluators. Support vector regression models dedicated to the periods when the speaker is speaking, the listener is speaking, and no one is speaking are built with non-verbal features extracted from facial expressions, head movements, prosody and speech turns. The resulted accuracy was not high but showed the potential of the proposed method.

Keywords: Active listening · Elderly support
Multimodal interaction

1 Introduction

The population of elderly people is growing rapidly in developed countries. If they do not maintain social life with others, they may feel loneliness and anxiety. For their mental health, it is reported effective to keep their social relationship with others, for example, the conversation with their caregivers or other elderly people. There are already some non-profit organizations recruiting volunteers for engaging "active listening" with the elderly. Active listening is a communication technique that the listener listens to the speaker carefully and attentively. The listener also ask questions for confirming or showing his/her concern about what

G. Meiselwitz (Ed.): SCSM 2018, LNCS 10914, pp. 330–341, 2018.
https://doi.org/10.1007/978-3-319-91485-5_25

the speaker said. This kind of support helps to make the elderly feel cared and to relieve their anxiety and loneliness. However, due to the lack of the number of volunteers comparing to that of the elderly who are living alone, the volunteers may not be always available when they are needed. In order to improve the results, always-available and trustable conversational partners in sufficient number are demanded.

The ultimate goal of this study is the development of a computer graphics animated virtual listener who can engage active listening to serve elderly users at a level close to human listeners. In order to conduct successful active listening, it is considered essential for the listener to establish the rapport from the speaker (elderly users). Rapport is a mood which a person feels the connection and harmony with another person when (s)he is engaged in a pleasant relationship with him/her, and it helps to keep long-term relationships [8,10]. In order to achieve this, like a human listener, the virtual listener has to observe the speaker's behaviors, to estimate how well the speaker is engaging the conversation [15], and then decides how it should respond to the user.

The estimation of the level of the speaker's engagement in the active listening conversation is therefore one of the essential functions of the active listener agent. In the context of active listening, the level of the speaker's engagement in the conversation can be considered to be expressed by his/her attitude toward the listener. The utterances of the speaker are obvious cues for the estimation of the speaker's engagement. However, due to the nature of active listening conversation, the speaker may utter in arbitrary contexts, It is difficult to utilize this information. Non-verbal behaviors are considered more general and more robust (less user-dependent). On the other hand, benefits from the advance of sensor device technology, machine learning models from multimodal sensory information has been proving to be effective in estimation human communication behaviors [2].

This paper reports our progress in developing the estimation model of speaker's engagement for active listener agents with non-verbal features based on a multimodal corpus of active listening conversation. Since there are only three situations in dyadic active listening, the speaker is speaking, the listener is speaking, neither of the two participants are speaking, we developed the model for each situation because the available features are different in each situation. The non-verbal features include head movements, facial expressions, prosodic information, and speech turn information.

2 Related Works

The research works on making robots and agents to be the communication partners of the elderly and dementia patients have been getting popularity. One of the methods to mitigate the progression of dementia is "coimagination" which was proposed by Otake et al. [13]. It is a method by using pictures as the references for the topics in group conversation. All participants who are elderly people have equal chance to listen, to talk, to ask questions, and to answer questions. It is reported that the elderly who participated in this activity talked and smiled

more frequently than before. However, this method has the limitation that all of the participants have to meet at one single place which may be difficult in practical.

Bickmore et al. [3] proposed a companion agent to ease the anxiousness of elderly inpatients. Huang et al. [10] developed a rapport agent which reacts to facial expressions, backchannel feedbacks, and eye gazes of the user. The agent is designed to show behaviors which are supposed to elicit rapport. However, it does not try to estimate and react to the user's internal state or "mood". For example, when the user looks in bad mood, showing the agent's concern on the user by saying "Are you OK?" like human do. The SEMAINE project [12,14] was launched to build a Sensitive Artificial Listener (SAL). SAL is a multimodal dialogue system with the social interaction skills needed for a sustained conversation with the user. They focused on realizing "really natural language processing [4] " which aims to allow users to talk with machines as they would talk with other people.

These works were developed base on the subject studies in the U.S. or in other western countries where the subjects' communication style may diverge from that of Japanese ones [7]. In this study, we collected an active listening corpus of Japanese subjects and analyzed Japanese style verbal/non-verbal behaviors which potentially improve the effectiveness of a listener.

3 Active Listening Corpus

This work shares the same data corpus with our previous work [9]. The corpus is collected in a human-human teleconferencing experiment which is conducted to imitate the situation where a virtual listener which is displayed on a 2D surface. The collected video data were evaluated and annotated in the aspect of how positive the speaker's engagement attitude by both of the participants and a third person from an objective viewpoint. The annotated scenes are then extracted for the development of automatic estimation method.

3.1 Experiment Setup

Due to the difficulty in recruiting elderly participants, and the experiment procedure explained in the following sections may be difficult the elderly, college students are recruited in this experiment. Furthermore, we considered the implicit criteria in judging communication partner's attitude to be positive or negative should not vary largely between younger and senior generations.

Eight pairs of participants in the same gender were recruited as the experiment participants. All of them were college students and native Japanese speakers. The two participants of each pair were recruited with the condition that they are close friends. This is because close friends were assumed easier to talk with each other in the limited experiment period. In order to simulate the situation of talking with a 2D graphical agent, the participants of each pair were separated

into two rooms (Figs. 1 and 2) and talked with each other via Skype teleconferencing software. In each session, one participant played the role of speaker, and the other one played the role of the listener.

They were instructed to sit on a chair so that the move of their lower bodies can be controlled within a limited range. Each room was equipped with two video cameras. One was used for recording the participant from the front. The other one was used for logging the Skype window which was duplicated on another monitor. The speaker talked with the listener who was projected on a large screen near life-size. The height of the projected image was adjusted so that the speaker can see the listener's eyes roughly at the level for eye contact. Natural head movements and eye gazes shifts can be further analyzed.

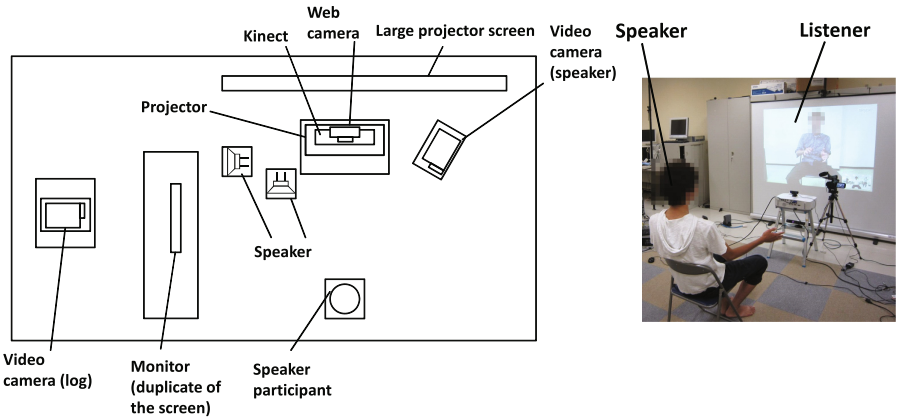

Fig. 1. Setup of the room where the speaker participant was in. The listener was projected roughly as life-size and the second monitor was used for video logging

3.2 Experiment Procedure

Speaker participant initiates the session and talks to the listener about his/her family. The topics of the conversation were "pleasant experience with family" and "unpleasant experience with family." These topics were chosen because they are common for almost everyone including the young experiment participants and the elderly. Listener participant was instructed to try their best to be a good active listener. That is, listen to the speaker carefully and attentively, follow the speaker's talk with questions or other feedbacks like nods or laugh as possible as the participant can.

They interchanged their roles in the sessions and started to talk from the pleasant experience at first because it should be easier to do (Table 1). The duration of one session was set to be seven minutes because it is considered long enough for the participants to start to talk something meaningful and keep the

Table 1. Topic and subject assignment of each active listening session

Session	Topic	Speaker	Listener
1	Pleasant experience with family	A	B
2		B	A
3	Unpleasant experience with family	A	B
4		B	A

whole experiment within a reasonable time period. During the experiment, the experimenters were outside of the two rooms without intervening the participants.

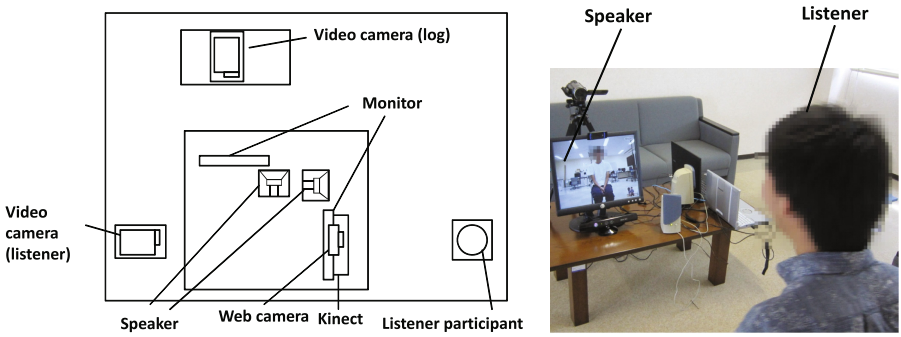

Fig. 2. Setup of the room where the listener participant was in. The second monitor was used for video logging

3.3 Evaluation of Speaker's Attitude

After the experiment, four evaluators (two male college students and two female college students) were recruited to evaluate the attitude of the participants by annotating on the recorded video corpus. The recruitment was done in the condition that the evaluators neither participated the active listening experiment, nor have close relationship with the participants to ensure objective evaluation results. How the attitude of the speaker is supposed to be perceived by the listener, and how he/she perceived the listener's attitude were evaluated in 7-scale measure from value 1 (negative) to 7 (positive). The evaluators were instructed to evaluate the participants's conversation base on their observation and perception on both verbal and nonverbal behaviors of them. Appropriate back-channel feedbacks like nods, questions, silence, agreeing opinions, smiles, or laughs were provided as positive examples in the instruction to the evaluators. In order to prevent gender bias, the video data of each participant pair were annotated by one male evaluator and one female evaluator. The video annotation tool,

ELAN [11] was adopted for this purpose. In order to align the granularity and label positions among different evaluators, they were instructed to annotate their evaluation with the four following rules:

1. The whole time line has to be labeled without blank segments
2. Starting and ending positions of the labels should be aligned to utterance boundaries
3. One label can extend to multiple utterances
4. The duration of one individual label is at most 10 s

Figure 3 shows an example segment after the evaluation annotation. Although the label values are discrete, the label values are immediately successive to one another, that is, there is always a label value at any time point.

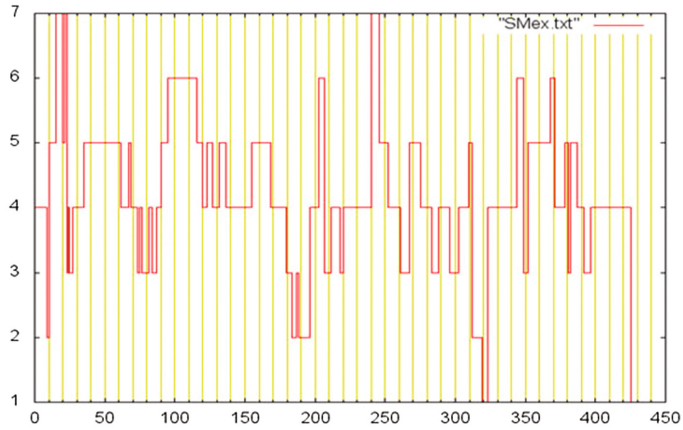

Fig. 3. Conceptual diagram of evaluation value sampling. The horizontal axis shows time (ms) and the vertical axis shows evaluation values

After the evaluation annotation, totally 3,644 labels were gathered from the four evaluators. Table 2 shows the number of labels in each value. From the distribution of the data values, varieties can be observed among the evaluators.

Table 2. Number of the 1 to 7 value labels gathered from all four evaluators

Evaluator	1	2	3	4	5	6	7
A	15	39	80	235	253	88	46
B	26	50	92	188	345	55	23
C	37	98	193	413	276	78	34
D	22	54	113	302	284	138	67

4 Automatic Estimation of Speaker's Engagement

4.1 Estimation Target Values

In order to realize automatic estimation of speaker's engagement attitude, there are two issues have to be solved. The first one is when to generate estimation results, and the second one is what should be the target value of the estimation. For a virtual listener agent to function properly, it has to show its behaviors to react to speaker's behaviors in reasonably short intervals to allow the speaker to feel that it is responsive and life-like. The period of one utterance which is usually within several seconds serves to be a candidate. In a dyadic conversation, there are three possible situations regarding to utterances, the speaker is speaking, the listener is speaking, and no one is speaking. Since the available information are not equal in these situations, they are estimated by dedicated models that we call *speaker*, *listener*, and *silences*, respectively. The estimated speaker's engagement attitude are generated at the end of each period by the corresponding model.

During the evaluation process, the measurement was subjective. A particular value of the annotation may mean different degree of engagement of the speaker to individual evaluators. In order to sum up the results from different evaluators, the elimination of the bias caused by the evaluators is required. The label values are then standardized with Z score regarding to individual evaluators. Table 3 shows these normalized label values. From the observation on the table, the average (Z score: 0) of the evaluators can be found to tend to be higher than the middle value (4) of the 1 to 7 scale except evaluator C.

Table 3. Z score normalized label values of each evaluator

Evaluator	1	2	3	4	5	6	7
A	−0.781	−0.562	−0.343	−0.124	0.096	0.315	0.534
B	−0.802	−0.603	−0.405	−0.207	−0.009	0.190	0.388
C	−0.750	−0.499	−0.249	0.002	0.252	0.503	0.753
D	−0.772	−0.544	−0.316	−0.089	0.140	0.367	0.595

The original annotation contained only the integer values between 1 and 7. After the Z score transformation, the label values were transformed to various real number values roughly ranging from −0.8 to 0.8. Target values of automatic estimation are then set to be the average values of the two evaluators annotation label values. Since the label boundaries are not aligned to the time periods of automatic estimation, label values are weighted regarding to time intervals. The data of 13 participants (nine were male and four were female, 26 experiment sessions) were used for the model training. Table 4 shows the distribution of the target values of automatic estimation for each model. They are used in the machine learning phase, one individual period is used as one instance for training or validation.

Table 4. Distribution of the data sets used in the training of speaker, listener, silence models

	Instances	Num per session	Max	Min	Average	Stdev
Speaker	3,099	119.2	0.844	−0.775	−0.020	0.226
Listener	524	20.2	0.662	−0.737	−0.013	0.213
Silence	3,648	140.3	0.844	−0.775	−0.019	0.226

4.2 Feature Sets

The speaker's behaviors during the duration of individual label were considered to affect the judgement of the evaluators and were used as the cues for the estimation of the speaker's attitude (or the level of engagement in the conversation with the listener). All behaviors including verbal ones and nonverbal ones of the speaker may affect the evaluators' perception. Regarding to the evaluation of active listening conversation, we selected the low-level communicational facial expressions, head movements, prosody and speech turn for the estimation. These features were selected base on the following hypotheses. Smiles may imply that the speaker is in his/her pleasant mood. Nods may imply that the speaker agrees to what the listener is talking about or shows his/her willing to listen to the listener. The speaking frequency of the speaker may imply his/her willing to talk with the listener. Prosodic information in the voice of speakers is also considered to propagate the emotional state of the speaker. The followings are the list of all 98 features used in the estimation.

- Facial expression (F): features related to facial expressions. Most features are extracted with face recognition software tool, visage|SDK[1] at 30 fps. Action units of facial expressions are extracted according to CANDIDE model [1] which is designed for face tracking rather than well-known FACS [5] in psychology field (59 features).
  - Average intensity of the following action units in current estimation period: nose winker, jaw push z/x, jaw drop, upper lip raiser, lip stretcher left/right, lip corner depressor, left/right outer brow raiser, left/right inner brow raiser, left/right brow lowerer, left/right eye closed, rotate eyes left, rotate eyes down, and lower lip x push
  - Intensity values of the same action unit set immediately before the estimation (average of the frames in last 10 ms)
  - average intensity of the same action unit set up to now
  - number of smiles per second in the current estimation period (hand labeled)
  - number of laughs per second in the current estimation period (hand labeled)
- Head movements (H): features related to head movements. Most features extracted with visage|SDK at 30 fps (19 features).

[1] http://visagetechnologies.com/products-and-services/visagesdk/.

  - Average three-dimension head position (x, y, z) of current utterance
  - Average three-axis head rotation ($roll$, $pitch$, yaw) of current utterance
  - Three-dimension head position immediately before the estimation (average of the frames in last 10 ms)
  - Three-axis head rotation immediately before the estimation (average of the frames in last 10 ms)
  - Average three-dimension head position up to now
  - Average three-axis head rotation up to now
  - Number of nods per second during current utterance (hand labeled)
- Prosody (P): the prosodic information when the speaker-role participant is speaking. Praat[2] was used to compute the following prosodic features of the the current utterances. Since the raw values can have a large diversity among the speakers, relative values are used here (11 features).
  - Ratio of average pitch/intensity of current utterance comparing to the average pitch up to now
  - Maximum pitch/intensity of current utterance comparing to the average of current utterance
  - Maximum pitch/intensity of current utterance comparing to the average of all utterances up to now
  - Minimum pitch/intensity of current utterance comparing to the average of current utterance
  - Minimum pitch/intensity of current utterance comparing to the average of all utterances up to now
  - Speech rate of current utterance
- Speech turn (T): the features related to speech turns (nine features).
  - Ratio of the time period when no one was speaking (silent) regarding to the total session time up to now
  - Average number of silent periods regarding to the total session time up to now
  - Average duration of silent periods up to now
  - Ratio of the time period when the speaker was speaking (speaking) regarding to the total session time up to now
  - Average number of speaker's speaking periods regarding to the total session time up to now
  - Average duration of speaking periods
  - Duration of speaker's last utterance
  - Speaker of last utterance preceding this one
  - Three-gram pattern of the speaker of last three utterances

4.3 Support Vector Regression Models

Support vector regression models are trained for the three periods, speaker is speaking (speaker model), listener is speaking (listener model), neither speaker nor listener is speaking (silence model). The SMO Regression implementation

[2] http://www.fon.hum.uva.nl/praat/.

of Weka [6] toolkit is used in data mining experiments. All four feature sets are used in training speaker model while prosody feature set was absent in the training of listener and silence models. PUK kernel [16] was adopted because it achieves best results in the experiments based on our data corpus. The complexity parameter were explored from one to 10 (10 steps) and the combinations of normalization/standardization were explored to find optimal results. Correlation and R^2 (coefficient of determination) are used as the metrics in evaluation model performance. From the definitions, correlation is computed with equal weights regardless of data values while the errors at extreme values has larger impacts in the computation of R^2.

Figures 4 and 5 show the 10-fold cross validation results of each estimation model regarding to all combinations of feature sets. Training parameters are aligned to the values which achieve highest correlation with full feature set of each model (TPFH for speaker and TFH for listener/silence models). The order of the performance of the three models is silence, speaker, and listener. The fact that silence model performed best in spite of less available features may imply the influence of verbal information in the judgement of speaker's engagement in speaker and listener models. Generally but not always, better performance is achieved with more feature sets. Speaker model has best performance (correlation: 0.55, R^2: 0.35) with feature set TFH, listener model has best performance (correlation: 0.41, R^2: 0.18) with feature set H, silence model has best performance (correlation: 0.59, R^2: 0.39) with feature set FH. Facial expression features and head movement features have larger impact on the results while the prosodic information is least effective.

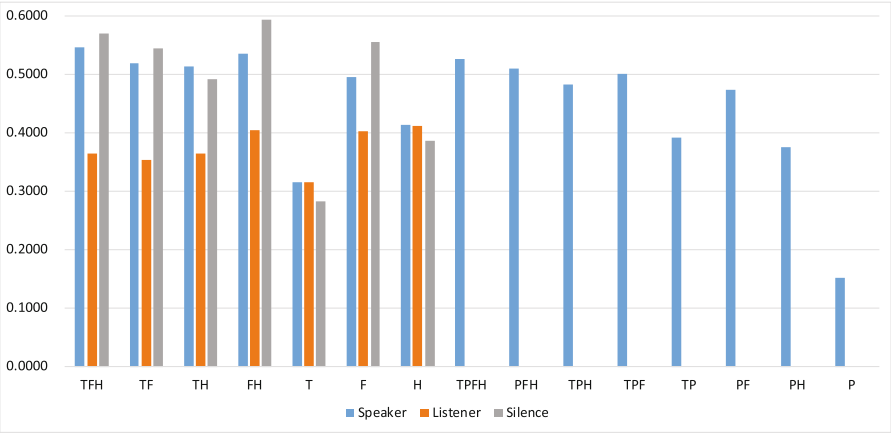

Fig. 4. Correlation results of the estimation models, speaker, listener, silence regarding to feature sets

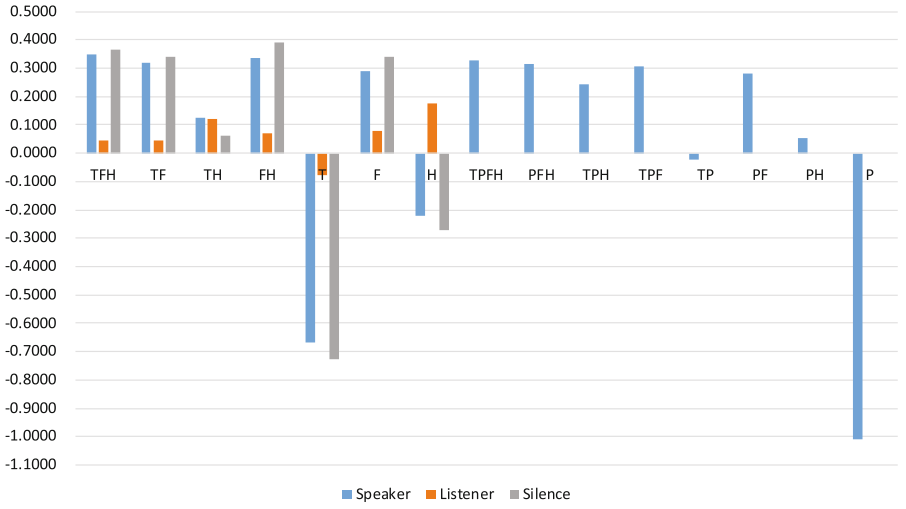

Fig. 5. R^2 results of the estimation models, speaker, listener, silence regarding to feature sets

5 Conclusion

In order to realize an active listener agent for the elderly, the ability for the agent to evaluate the engagement level in the conversation is essential. In this work, we conducted an active listening conversation experiment with human participants. The speakers' attitude was subjectively annotated into seven levels by human evaluators. In order to combine the results from all participants and evaluators, the label value stream was standardized to Z-score values regarding to the evaluators. Nonverbal features including facial expressions, head movements, prosody and speech turns are then used to train support vector regression models to estimate speaker's engagement attitude in three situations, the speaker himself/herself is speaking, the listener is speaking, no one is speaking. The results are not high yet but showed the potential to solve this estimation task with non-verbal multimodal features.

In the future, at first we would like to increase the corpus size with additional experiments to obtain more stable results. Other features like postures and gestures can be explored to improve the performance of the models. The behaviors of the elderly can be quite different to young peoples, we would like to collect the data corpus with older participants and see whether the features can perform well enough. When the technology becomes matured, we will incorporate this function to a fully working listener agent and evaluate it with the elderly for long period.

References

1. Ahlberg, J.: Candide-3 – an updated parameterised face. Technical report. LiTH-ISY-R-2326, Department of Electrical Engineering, Linkoping University, January 2001
2. Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: a survey and taxonomy. CoRR abs/1705.09406 (2017). http://arxiv.org/abs/1705.09406
3. Bickmore, T., Bukhari, L., Vardoulakis, L.P., Paasche-Orlow, M., Shanahan, C.: Hospital buddy: a persistent emotional support companion agent for hospital patients. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS (LNAI), vol. 7502, pp. 492–495. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33197-8_56
4. Cowie, R., Schröder, M.: Piecing together the emotion jigsaw. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 305–317. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-30568-2_26
5. Ekman, P., Friesen, W.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1978)
6. Frank, E., Hall, M.A., Witten, I.H.: The weka workbench. In: Data Mining: Practical Machine Learning Tools and Techniques, 4th edn. Morgan Kaufmann (2016)
7. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind, 3rd edn. McGraw-Hill, New York (2010)
8. Huang, H.H., Matsushita, H., Kawagoe, K., Sakai, Y., Nonaka, Y., Nakano, Y., Yasuda, K.: Toward a memory assistant companion for the individuals with mild memory impairment. In: 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC 2012), Kyoto, pp. 295–299, August 2012
9. Huang, H.H., Shibusawa, S., Hayashi, Y., Kawagoe, K.: Toward a virtual companion for the elderly: an investigation on the interaction between the attitude and mood of the participants during active listening. In: 1st International Conference on Human-Agent Interaction (iHAI 2013), Sapporo, Japan, August 2013
10. 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 (LNAI), vol. 6895, pp. 68–79. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23974-8_8
11. Lausberg, H., Sloetjes, H.: Coding gestural behavior with the NEUROGES-ELAN system. Behav. Res. Methods **41**(3), 841–849 (2009)
12. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: IEEE International Conference Multimedia and Expo, pp. 1079–1084 (2011)
13. Otake, M., Kato, M., Takagi, T., Asama, H.: Coimagination method: supporting interactive conversation for activation of episodic memory, division of attention, planning function and its evaluation via conversation interactivity measuring method. In: International Symposium on Early Detection and Rehabilitation Technology of Dementia, pp. 167–170 (2009)
14. Pammi, S., Schro, M.: Annotating meaning of listener vocalizations for speech synthesis. In: 3rd International Conference on Affective Computing and Intelligent Interaction (ACII 2009), pp. 1–6 (2009)
15. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychol. Inq. **1**(4), 285–293 (1990)
16. Ustun, B., Melssen, W., Buydens, L.: Facilitating the application of support vector regression by using a universal pearson vii function based kernel. Chemometr. Intell. Lab. Syst. **81**(1), 29–40 (2006)