Word guessing game with a social robotic head

Štefan Beňuš[1,2], Róbert Sabo[2], Marián Trnka[2]
[1] Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 94901 Nitra, Slovakia
[2]Institute of Informatics of the Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovakia
[1]sbenus@ukf.sk, [2]robert.sabo@savba.sk, [2]trnka@savba.sk

Abstract. *In this paper we address three limitations of our previous implementations in human-machine spoken interaction in Slovak: low prosodic variability, limited naturalness, and deployment in real acoustically challenging situations. We designed a word-guessing game in which subjects provide verbal cues for a target animal and social robotic head Furhat guesses the animal. We then deployed it in an open-air science festival with over 60 subjects playing the game. We describe the implementation and initial observations from the design, deployment, and user evaluation.*

1. Introduction

Spoken interactions between humans and machines (HMI) are becoming ever more common in everyday lives. Predictions for near future include common deployment of not only personal assistants in smart phones but also companions for social well-being for senior citizens, applications participating in diagnosing health issues, or teaming between humans and robots for various tasks. While most of the research in this area is done in English or other major languages, understanding both cognitive as well as engineering aspects of deployment in less studies languages is also warranted. In this paper we address limitations of our previous implementations in human-machine spoken interactions in Slovak and describe the implementation and initial observations from a novel research tool for spoken interaction between humans and robots. Our long-term goal is to understand better the intricacies and challenges of HMI in Slovak and inform thus designers developing real-world applications in this area.

In previous experiments with spoken HMI in our lab we used a simple card game GoFish, with a closed set of questions the user could ask the machine, or one-person adventure game motivated by Harry Potter, with a less constrained options for user to respond to the prompts from the machine. Our experience pointed to three areas of limitations relevant to the current study. First, when humans interacted with spoken dialogues systems in these task-oriented game-like scenarios ([1], [2], [3]) they tended to use speech with limited prosodic variability and engagement. We hypothesized that the absence of emotional attachment and social contact prior to the target dialogues might negatively affect both prosodic variability of humans and potential for speech entrainment between the humans and the machines ([2], [4]).

Second, the limited naturalness of the dialogues in our previous scenarios could also be attributed to the lack of backchanneling and conversational fillers that are so ubiquitous in human-human dialogues but were missing in our previous implementations.

Third, the experiments were conducted in laboratory environment without testing the deployment capability in real environments and situations and in adverse acoustic conditions. Transferring HMI applications from the laboratory to real environment and testing its usability is important if spoken HMI technology should be usable in a wide scale of users and situations.

All of these potential limitations are addressed in our new experimental setup in which we designed a simple games 'Guess my animal' and tested how people interact with the socially expressive robotic head Furhat with implemented backchanneling behavior in real acoustically challenging condition of open-air science festival with high babble noise and loud-speaker noise. Additionally, we tested the relevance of small talk prior to the target interaction, which was hypothesized to facilitate the establishment positive social contact, engagement of the users and perceived naturalness of the interaction.

2. Methods

We designed a simple game 'Guess an animal' in which a user selects a card with a name of the animal, provides cues and information about the animal without reveling its identity, and a machine guesses the animal. We employed the robotic head Furhat equipped with Slovak TTS and ASR and deployed in an open-air science festival with over 60 subjects playing the game. In this section we describe in turn the hardware, the robotic head and its setup, the software in terms of ASR, TTS and the implementation of the game, and the procedure for data collection.

2.1. Hardware

2.1.1. Furhat robotic head

The robotic head FURHAT is a 3D humanoid agent that employs the optical projection of an animated facial model ([5], furhatrobotics.com). The neck uses two degrees of freedom, which enables simple gestures like nodding or shaking and the movement of the head into any direction within a reasonable viewing angle. The flexibility of the facial animation model, allowing for gaze changes, eyebrow movement, blinking, syncing lip movements with speech, or various emotional gestures like disgust or happiness, together with the neck flexibility enable FURHAT to participate in social spoken interactions with humans and signal various intents and behaviors.

In terms of hardware, FURHAT is a computer with a mounted model of human head, the face is back-projected on the front mask; see Fig 2. In our experiment we used the 1st generation of FURHAT [https://docs.furhat.io/gen1]. Camera or Microsoft Kinect sensor using face recognition allows tracking the face of the user(s) who enter or remain in the Furhat's visual field and subsequent controlling of the head movement and gaze for eye contact functionality.

We used the default artificial face 'Bertil' and the control of the visual modality was kept at a minimum in the current experiment.

FURHAT comes pre-equipped with various software features. Primarily, it uses commercially available speech recognition (ASR), e.g. by Google, and speech synthesis (TTS), e.g. by Amazon. Our long-term goal is to use Slovak ASR and TTS in experiments investigating how modifications of TTS and improvements in ASR affect user experience. For this reason, although Slovak ASR is available with Google, we equip FURHAT with our own Slovak TTS and ASR so that we have greater control over the characteristics and deployment of Slovak ASR and TTS.

The robot can be programmed using FURHAT Legacy SDK [http://www.furhat.io] statechart XML-based framework for developing multi-modal interactive systems developed from the original IrisTK system [6]. This SDK is especially designed to be a powerful framework for social robotics applications. In our initial effort, however, we chose to control FURHAT over TCP/IP using Broker functionality [https://docs.furhat.io/gen1/tutorials/tutorial_broker_cshar]. This solution allows us to develop the game in any programming language and for the time being to circumvent technical problems in implementing our own Slovak ASR onto FURHAT.

2.1.2. *Hardware setup*

In the current experiment, the architecture of the hardware is illustrated in Fig. 1. The human user is sitting in front of the robot at a distance of about one meter and his/her speech is recorded with a head mounted microphone that is connected to a laptop. The recorded sound is sent to the ASR server that outputs the text transcript of the recognized utterances. The game itself and the dialogue manager (see section 2.2.3) processes the ASR output and selects the robot's behavior consisting of speech, non-verbal audio expression like hesitation or a backchannel, visual expression of the face, or a combination of the above.

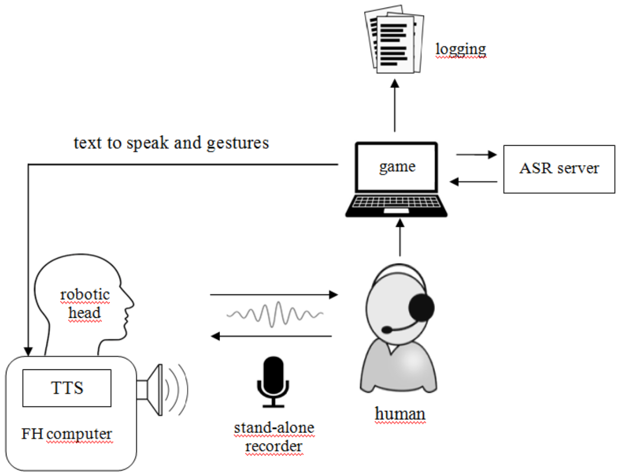

Fig. 1. Schematized experimental setup

2.2. Software

2.2.1. *Slovak ASR*

Slovak ASR was implemented based on Kaldi toolkit for speech recognition [7]. We used a generic language model trained on newspaper database collected from internet with 550k vocabulary items [8] and the acoustic model trained on TV news [9], [10]. The advantage of speech in TV news is that it contains spontaneous speech and background noise similar to our experiment. For this first experiment the acoustic and language models were not adapted in any way to the game domain. To make communication with the robot as natural as possible and to minimize the time delay, the recognizer runs in online mode. Hence, speech was recognized continually.

2.2.2. *Slovak TTS*

To make speech synthesis fast we used a statistical text-to-speech system wherein speech is synthesized directly from hidden Markov models [11], [12]. The statistical model for Slovak female voice is trained on a newly created phonetically balanced speech corpora consisting of 10k sentences spoken by professional actor. The TTS was implemented in FURHAT using Microsoft Speech API 5.4 that supports syncing lip movements with audio through 21 visemes that were mapped to the Slovak ones [13].

2.2.3. *Game design and implementation*

The game is construed as a research tool for investigating various aspects of human-machine spoken interactions that can be deployed and adjusted to various situations. For this reason, in this first implementation we aimed at a simple algorithm to test the functionality of the basic building blocks rather than a sophisticated and complex dialogue manager.

The basis of the game is the table that lists the animals and describes their ***keywords***. There were 12 animals in total in this version of the game. We have grouped the keywords into seven ***attribute*** categories: {legs, food, area, species, characteristics, look, alternative}. We have pre-tested with several naïve subjects and added the keywords used by the pilot subjects. A subset of the table relevant for 'tiger' is illustrated in Table I.

Table 1. Attributes and keywords for the animal "Tiger"

Attribute	Keyword
Legs	four, four-legged
Food	meat, carnivore
Area	Asia Asian India Indian Bangladesh Russia
Species	Mammal vertebrae felidae
Characteristic	beast cat feline hunter hunt predator forest prairie jungle rainforest taiga teeth endangered
Look	stripe black brown orange white large
Alternative	Leopard

The ASR output is fed to the lemmatization process for Slovak [14] since Slovak has a rich system of inflectional morphology. The algorithm then searches for keywords representing the attributes in the lemmatized ASR output. If a relevant keyword is recognized, the total score for each animal with such a keyword is increased. Commonly, a keyword is applicable to several animals, e.g. 'big teeth' can describe both the tiger and the crocodile, but some attributes are assumed to be unique to a single animal, e.g. 'meow' should be used only with the cat. After pre-testing we have also included ***special keywords*** such as reptile, mammal, amphibian, bird, insect, carnivore, four-legged, double-legged, bark, meow, wing, fly, ride, etc. These form a subset of the keywords such that the animals including these keywords will add an extra point but the animals without these keywords will have one point deducted from their score.

The progress from the start to the FURHAT's guess is controlled with several ***internal parameters***. First, the player must use a minimal number of words. If this limit is not reached and FURHAT detects silence, it prompts the player with general cues like 'Tell me something more' or 'What else?'. Second, a minimal number of questions from FURHAT has to be asked. In these questions FURHAT determines the animal with the highest score of the keywords as the most probable guess, checks the attributes for which it detected no keywords, and asks randomly about one of these attributes. For example, 'Tell me how many legs this animal has' or 'Where does this animal live'. Third, FURHAT tries to achieve a minimal difference between the top animal with the highest score and the next best one. The interaction of these three parameters keeps the game to a manageable length, and guarantees dialogue-like turn-exchanges and sufficient input from the player.

FURHAT then proceeds to the guess, which is either unique, if a single animal reaches the highest score or includes alternatives if multiple animals reached the top score FURHAT checks with the player if its guess was correct and based on ASR of the response appends the count of the correct responses.

In addition to the internal parameters, two ***external parameters*** were also implemented. First, to test the effect of positive social and emotional engagement between the layer and the robot on the subsequent spoken interaction, we designed introductory ***small-talk***. Prior to playing the game, the user can interact with FURHAT in a short mini-dialogue. It consists of (1) FURHAT introduction and the prompt for the user's name, commenting on the beauty of the name, and if the name is correctly recognized, addressing the user with this name throughout the game, (2) FURHAT's asking about the user's birthplace, using it in its response if correctly recognized, and commenting positively about the place, and (3) asking about recent ice-hockey world championship, the preferred team, and again expressing positive comment regarding that team. We hypothesized that this small talk establishes positive emotion and social rapport of the user toward FURHAT since it expressed positive comments about user's name, birthplace, and favorite team. The interaction could thus be initiated with this small-talk or without it directly proceeding to the game.

Second, to enable testing the effect of ***backchannel*** use on the naturalness of the interaction, we selected four instances of backchannel 'mhm' from the corpus used to train the female TTS voice (section 2.2.2). If a short silence was detected in the running ASR, FURHAT produced one of these backchannels randomly and simultaneously nodded its head. We hypothesized that this behavior will naturally prompt for more input by the user and would be unobtrusive even if resulting in simultaneous speech of the user and the FURHAT. In the current setup, this parameter was always set to be active.

2.3. Procedure

If a passer-by was interested in interacting with FURHAT and FURHAT was not engaged with another player, the experimenter seated the subject in front of the FURHAT and fitted him/her with a head-mounted microphone. After briefly explaining the rationale of the game, the experimented started the external recorder and prompted the subject to test the game with the experimenter. The goal was not only for the subject to try out and get comfortable with giving cues about animals, but also to obtain the ***baseline*** of the user's speaking behavior. For later experiments, we plan to use this baseline in testing how the speaking behavior changes when interacting with an experimenter and with FURHAT.

After this trial run, the experimenter explained that the subject will play the game three times with FURHAT, that the experimenter will not participate in the game, and prompted the subject to use full sentences and rich input for the FURHAT. Importantly, motivation of a small gift was indicated if FURHAT is successful at guessing the animals.

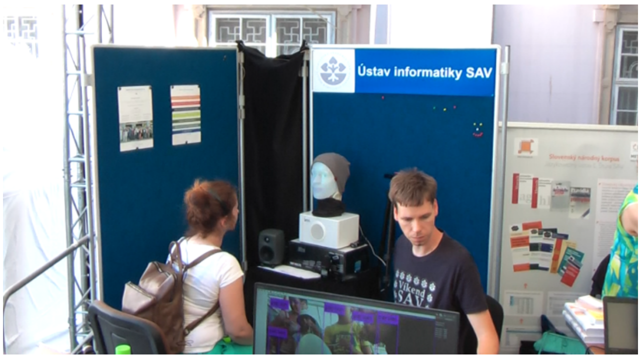

Fig. 2. Playing the game "Guess my animal" at the open-air science festival “Weekend with the Slovak Academy of Sciences”.

Before each game the subject selected a card with the target animal from a list of cards offered by the experimenter.

After completing the three rounds of the game, the experimenter asked the subject to fill out a brief post-test questionnaire primarily assessing two domains on a Likert scale from 1 (positive) to 5 (negative): FURHAT's speaking behavior (natural – not natural) and FURHAT's abilities (smart – dumb).

3. Results and observations

We start with the only quantifiable result available at the moment regarding the effect of small talk on the interactions and continue with informal observations regarding the functionality of the proposed set-up.

Our logs show that there were 73 subjects that completed the full three rounds of the game and we have 61 filled out post-test questionnaires. We calculated the total word and syllable count in all the recognized speech of the subjects from the logs. We take this as a proxy measure for user's engagement in the interaction with FURHAT. Additionally, we have the subjective evaluations present in the post-test questionnaire.

Of the 73 complete logs, 45 subjects started with the small-talk and 28 proceeded directly to the game. Welch two-sample t-test showed that interactions with small-talk included significantly more speech from the player than those without it; t[69.95] = -2.93; p = 0.0046 for syllables and t[69.37] = -2.65; p = 0.01 for words. This is shown in Fig. 3 and suggests that small-talk has positive effect on players' engagement with FURHAT in this implicit measure.

Of the 61 post-test questionnaires, 43 subjects played with small-talk. We observed a tendency that their experience with FURHAT's speech was slightly less natural than for the subjects without small-talk (t[30.4] = -1.77, p = 0.087). It might be that small-talk affects differently the explicit perceived naturalness of FURHAT's speech and the implicit measure of engagement. When people establish greater social rapport with FURHAT in small-talk, it increases their expectation of the naturalness of its speaking behavior, which is not always met in the current implementation. This speculation is supported by the results from the effect of small-talk on the perceived abilities of FURHAT. Subjects with small-talk preceding the game rated FURHAT as significantly less capable and smart than those without small-talk (t[43.9] = -2.63, p = 0.012).

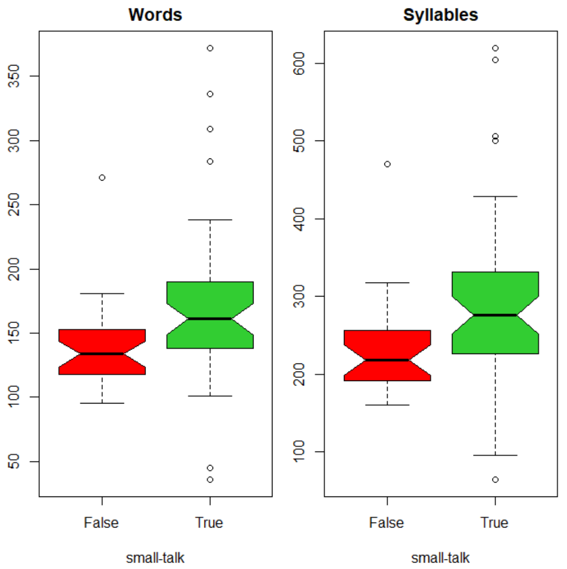

Fig. 3. The effect of small-talk on the number of words and syllables per subject

The experience with a 2-day deployment of FURHAT with Guess my animal game in an open-air science festival yielded the following observations. First, we discuss the limitations that we wanted to address. The implementation of the game with FURHAT was perceived very positively and the interaction as quite natural. This was the case both in the questionnaires (mean perceived speech naturalness 1.87 and abilities 1.71 on a scale between 1 and 5), and informal reactions of the subjects. The implementation of small-talk proved to affect engagement of people and their evaluation of the interaction. The speaking behavior with backchaneling and conversational fillers was also perceived as very natural. Finally, the fact that the game was functional in such adverse noise conditions (speech from passer-bys, loud music, loud presentations for big audience) is encouraging for future implementations of FURHAT with general public.

Additional observations relevant for future work include:

- The temporal management of the turn-taking should be improved as this was the only feature commented on by the subjects sometimes in negative terms. The main reason for this deficiency was that sometimes FURHAT's speech served as the input to the ASR, which resulted in asynchrony between the subject and FURHAT. This stems from the absence of information regarding the end of the TTS utterance sent to FURHAT. We will implement Voice-activity-detection and experiment with FURHAT's broker functionality to address this issue.
- The audio output from the speakers was more natural than using headphones but in high noise people sometimes did not hear FURHAT very well. The use of more powerful speaker system for open-air events is advisable.
- The recorded baseline when subjects interacted with the experimenter is not usable in the current setup due to great noise of the environment and the difference between the game external and internal speech recording. Adjustments should be made to use either directional microphone or include the pre-test baseline interaction into the game to unify the recording of entire session.
- The implementation of backchannels and conversational fillers was well received but some deployments were not natural. Particularly, backchannels after questions from the subjects or subjects utterance such as 'I don't know'.
- In this initial setup, the three renditions of the game played in succession did not vary basic utterances of FURHAT; especially those initiating and concluding the game. For greater naturalness we will vary these utterances to limit the perception of 'mechanical' speech by FURHAT.
- Several people also commented on the confusing biological sex of FURHAT. TTS was a female voice, the facial features were gender-neutral, but in Slovak morphology the name 'FURHAT' is associated with the male gender and in some of FURHAT's utterances male gender in self-address was also used. We will implement interactions with either male or female consistent persona for FURHAT in voice quality, visual representation of the face, and speaking behavior.

Acknowledgment

This work was funded by the Slovak Scientific Grant Agency VEGA "Automatic assessment of acute stress from speech", grant No. 2/0161/18.

References

1. R. Levitan, Š. Beňuš, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Proc. Of Interspeech 2016*, 1166–1170.
2. Š. Beňuš, M. Trnka, E. Kuric, L. Matrák, A. Gravano and J. Hirschberg, "Prosodic entrainment and trust in human-computer interaction," in *Proc. of 9th International Conference on Speech Prosody*, pp. 220-224, 2018.
3. S. Beňuš, M. Patacchiola, M. Trnka, D. Zanatto, R. Sabo, A. Cangelosi, "Do people trust robots whose prosody synchronizes with the user?" in *Šašinka, Č.,*

Strnadová, A:, Šmideková, Z., Juřík, V. (eds.). Kognice a umělý život, sborník příspěvků [Cognition and Artificial Intelligence, conference proceedings], pp. 9-10, Brno: Flow, 2018.

4. M. Lohani, C. Stokes, M. McCoy, C. A. Bailey, and S. E. Rivers, ”Social interaction moderates human-robot trust-reliance relationship and improves stress coping,” *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 471-472, 2016.
5. S. Al Moubayed, J. Beskow, G. Skantze, B. Granström, ”Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction,” in *Esposito A., Esposito A.M., Vinciarelli A., Hoffmann R., Müller V.C. (eds) Cognitive Behavioural Systems.* Lecture Notes in Computer Science, vol 7403. Springer, Berlin, Heidelberg, 2012.
6. G. Skantze and S. Al Moubayed, “IrisTK: a statechart-based toolkit for multi-party face-to-face interaction,” in *Proceedings of ICMI.* Santa Monica, CA, 2012.
7. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. “The kaldi speech recognition toolkit” in *Proc. ASRU*, pp. 1–4, 2011.
8. M. Lojka, P. Viszlay, J. Staš, D. Hládek, J. Juhár, “Slovak Broadcast News Speech Recognition and Transcription System,” in: *Barolli L., Kryvinska N., Enokido T., Takizawa M. (eds) Advances in Network-Based Information Systems.* NBiS 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 22. Springer, Cham, 2019.
9. P.Viszlay, J. Staš, T. Koctúr, M. Lojka, J. Juhár, “An extension of the Slovak broadcast news corpus based on semi-automatic annotation,” in *Proc. of the 10th edition of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, pp. 4684-4687, 2016.
10. M. Pleva and J. Juhar, “TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation,” in: *Proc. of LREC2014*, Reykjavik, Iceland, ELRA, 2014, pp. 1709–1713
11. M. Rusko, M. Trnka, S. Darjaa and J. Hamar, “The dramatic piece reader for the blind and visually impaired,” in *Proceedings of SLPAT*, pp. 83-91 Grenoble, 2013.
12. M. Sulír, J. Juhár and M. Rusko,”Development of the Slovak HMM-based TTS system and evaluation of voices in respect to the used vocoding techniques,” in *Computing and Informatics*, vol. 35, no. 6, p. 1467-1490, 2016.
13. SPEECH API https://azure.microsoft.com/en-gb/resources/samples /cognitive-speech-tts/...
14. R. Garabík and M. Šimková, “Slovak Morphosyntactic Tagset,” *in Journal of Language Modeling.* Institute of Computer Science PAS, vol. 0, no. 1, pp. 41-63, 2012.