Calling um… John or *Calling John!*

The perceptual effect of prosody in voice-activated system responses

Eran Aharonson[1,4], Vered Aharonson[1,2], Talya Porat[3,4], Vered Silber-Varod[1,2]
[1]Afeka Tel Aviv Academic College of Engineering; [2]ACLP – Afeka Center for Language Processing; [3]Ben-Gurion University;
[4]ACMiX – Afeka Center for Mobile Intelligent Experience
{erana, vered, veredsv}@afeka.ac.il; talya@bgu.ac.il

***Abstract* – In this paper we try to improve the human-machine interaction of a voice-activated system by adding prosodic characteristics to the system. We focus on verbal hesitation, which is manifested by speech disfluencies. In human-human communication recent research shows that moderate disfluencies make speakers more credible. In addition, people tend to react more leniently to an erroneous answer, if the answer was given by the conversant in a hesitating manner, implying that the responding person is unsure of the correct answer. In this study we investigate the hypothesis that users will react in a similar way to voice activated systems. Specifically, we hypothesized that adding prosodic features to the system's speech responses, will increase the user's perception of the system credibility, his/her overall satisfaction and reduce frustration while using the system.**

Keywords: Multimodal Interaction; Human-Machine Interaction; prosody; speech recognition

INTRODUCTION

Humans are extremely sensitive to nuances in each other's intonation, tempo and other speech prosodic features. In human-human interaction, prosody helps dialogue partners to detect, correct and avoid communication failures. On the contrary, human-machine interactions, specifically between a user and a spoken dialogue system, exhibit relatively frequent communication breakdowns, due mainly to errors in the Automatic Speech Recognition (ASR) component of these systems. In [1], Pon-Barry and Shieber write: "while most people can think of an instance where they have interacted with a call-center dialogue system, or command-based smartphone application, few would argue that the experience was as natural or as efficient as conversing with another human. To build computer systems that can communicate with humans using natural language, we need to know more than just the words a person is saying; ….". In command-based applications, there are at least two important aspects that can improve the human-machine interaction. The first one is the *naturalness* of the interaction. This can be done by improving the robotic sound of the system and integrating to it prosodic features. The other aspect is the *transparency* of the interaction. The system has to reflect to the user its limitations and uncertainties. In a study on Uncertainty Visualization, [2] argue that most of the current applications, especially those in the realm of natural language processing, are statistically based, meaning they provide the "best guess" by the algorithm (based on training data, parameter settings and user input). This "best guess" output is just one of a very large collection of possibilities, however, the system presents only this single response without providing details about probabilities, uncertainties and the way the algorithm works. This lack of detail makes it easy to misconstrue the output as having a low uncertainty and prevents users to make well-informed decisions based on the reliability of the output.

In this study, we add speech disfluencies (pauses and filled pauses) to the system to model verbal hesitation, in order to increase the system's naturalness and transparency and thus improve system credibility and user satisfaction while decreasing frustration. Recent results support the conclusion that duration increase, achieved by the combined effects of pauses and retardations, is an acoustic correlate of hesitation ([3], [4]). Thus, disfluencies can both increase the *naturalness* of the synthesized speech [5] and its *transparency* – meaning they act as a paralinguistic signaling of uncertainty in the dialogue [4].

RESEARCH MODEL AND HYPOTHESES

Fig. 1 demonstrates the proposed research model and hypothesized relations between constituents. The description of the constituents, their relations and the research hypotheses are detailed below.

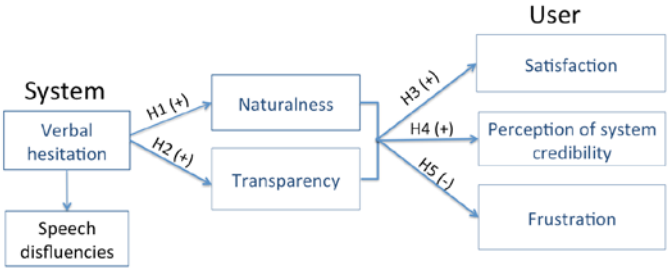

Figure 1. The proposed research model and hypothesized relations between constituents.

A. *Verbal Hesitation as an indication to naturalness of speech and system transparancy*

In human-human communication, when a person hesitates when answering a question, it usually implies that she or he is unsure of the correct answer. Pauses and retardations have been shown to be among the acoustic correlates of hesitations [3].

Therefore, we posit that speech disfluencies produced by people is a natural modality for a system to communicate uncertainty. Thus, adding similar human verbal hesitation (pauses and retardations) to the system will increase its naturalness.

According to [1], level of certainty is an important component of internal state. When people are interacting face

to face, they are able to sense whether the speaker is certain or uncertain through contextual, visual, and auditory cues. If we enable computers to do the same, we can improve the interaction between the voice-activated system and the user.

Thus, correlating the systems' hesitation level to the objective confidence level will enable the user to receive feedback about the system's inner state and it will become transparent. For example, when the user asks to "*Call Mary*", if the system recognizes the name "*Mary*" with a high confidence level, the system's response will be a confident one: "*Dialing Mary*"; in a lower recognition confidence level, the response will be more hesitant: "*Dialing…[pause] Mary*" or "*Dialing uh Mary*".

Hence, adding natural human hesitation manifested by speech disfluencies to the system will increase the system's transparency and expose its uncertainties.

Thus, we hypothesize:

H1: speech disfluencies modeling human verbal hesitation will increase the naturalness of the voice-activated system.

H2: speech disfluencies modeling human verbal hesitation will increase the transparency of the voice-activated system.

B. Naturalness and transparancy as increasing user satisfaction and system credibility, while decreasing frustration

A recent study [6] examined phone survey interviewers. They found that the most successful interviewers (the ones who convinced respondents to stay on the line and answer questions) spoke moderately fast and paused occasionally, either silently or with filler uh or um. The lowest success rates got the interviewers who made no pauses at all. The researchers assume it is because they sound too scripted. Thus, speaking with a certain number of uhs and ums, may enhance the speaker's credibility, and perhaps the system's credibility.

In addition, empirical evidence shows that "tutors respond differently to students based on their perception of the certainness of a student turn" [7: 1840]. In particular, tutors correct or paraphrase student answers more frequently for answers containing signals of uncertainty than for answers without uncertainty [8]. Perhaps, users will be more tolerant, satisfied and less frustrated if the system will be hesitant when it is not sure, since the user can get prepared for an erroneous answer. [9] argue that listeners' recognition benefits from any delay before a word, whether it's a silent pause or a filled pause, because the delay "attunes the attention."

Thus, we hypothesize:

H3: Increase in the system's naturalness and transparency will increase user satisfaction.

H4: Increase in the system's naturalness and transparency will increase user perceived system credibility.

H5: Increase in the system's naturalness and transparency will decrease user frustration.

To conclude, we propose that by adding verbal hesitation to the voice-activated system, we increase the user perception of the naturalness and transparency of the system. A more natural and transparent voice-activated system will increase user satisfaction, increase the user perception of the system's credibility and decrease the user frustration when the system is wrong.

METHOD

A. Experiment outline

To check the above hypotheses, we used a scenario of name dialing in a simulated Hebrew voice activated system, where the subject says "*[name]*" and the recognition system responds with "*Dialing [name]*". Based on the recognition confidence of our system, the system's output was either a confident answer or one of two hesitating answers.

Three types of systems were compared: System I: a monotonous ("Indifferent") system, which responds with the same confident voice, with no relation to the real recognition confidence level (similar to nowadays systems). System S: a prosodic-based ("Sensitive") system, that responds with increasing amount of hesitation to a decreasing confidence level. System M: a "Misleading" system – responds with increasing hesitation amount to increasing confidence level (i.e. hesitates in cases where the response is correct, and confident in cases where the response is incorrect).

Each session consisted of the subject evaluating two of the three systems. The pair of systems in the session as well as the order of their presentation was varied between subjects, to eliminate precedence effect.

For each system, the subject had to say seven different names – three repetitions of each name in a randomized order, but which did not allow subsequent identical names. Following those twenty names dialing, the subject filled an evaluation form, before moving on to a similar procedure for the second system. A concluding comparative evaluation form concluded the session.

The user sat next to a smartphone demo on a regular PC computer. The user was requested to imagine that she or he is driving a car and has to make some phone calls. The interface was very simple (Fig. 2): the user had to press the microphone icon to say the desired name to call and then the system would give its output.

Figure 2. User interface of the testing system.

Tewnty-two subjects participated in the study. Each subject signed a consent form and filled a demographic questionnaire, with additional questions on whether he/she has a Smartphone and used a speech recognition system.

The probability of correct recognition was set to $P_{correct}$ = 0.6 (resulting in 8 misrecognitions out of the 20 names).

B. The system responses

The system had three types of responses:

A confident, straight-forward respond:

[mitkasher le Haim Cohen] *'Dialing Haim Cohen*' (1)

A minor hesitation respond, which has an elongated preposition [le] 'to' (761ms long) followed by a 658ms pause, which precedes the target proper name "*Haim Cohen*":

[mitkasher le: ... Haim Cohen] '*Dialing ... Haim Cohen*' (2)

An uncertain respond, which has an elongated prepostion, as in (2), but also an additional hesitation marker [e:m] '*um*' (1.43sec long):

[mitkasher le: e:m Haim Cohen] '*Dialing um Haim Cohen*' (3)

Durational features of the responses were also monitored since, as studied in [10], the time for completing an action of voice dialing is perceived differently in different paradigms, although the actual completion time was similar. The length of the "confident" respond was 1.5 seconds of verbal respond. The "minor hesitation" verbal respond consisted of 3.5 seconds, and the "major uncertain" verbal respond took 4.14 seconds.

RESULTS

The results support our hypotheses that the "sensitive" system (S) was favorably perceived by the participants.

A. Speech disfluencies as increasing system's naturalness and transparancy

Participants who evaluated systems S ("Sensitive") and M ("Misleading") noticed the added hesitation in the speech. The feedback on both systems was that the systems are sensitive and behave human-like. Participants used the terms: *"natural", "human", "not the regular mechanic voice"* and *"hesitating"* to describe the systems.

In addition, participants which preferred system S said that they learned to listen to the hesitation and be more prepared for errors: *"when the system hesitated, it gave me a clue that the system might be wrong, on the contrary, when it was confident, I knew it got it right"; "I like it that the system knows when it is unsure"; "The first system [Sensitive] is softer, and you know when it is going to make a mistake, the other one [Indifferent] is too self-confident, it shoots the answer no matter if it is right or wrong"*.

On the contrary, participants who evaluated system M ("Misleading") said that the system was *"confused"* and *"weird"* because *"it is hesitating when it is correct and confident when it makes mistakes"*.

An interesting result was that participants which evaluated both systems I ("Indifferent") and S ("Sensitive") felt that besides the "humanize" speech of system S there was more time to correct the system if it made a mistake compare to system I where there was no time to correct the system when it was wrong.

B. Satisfactory, credibility and frustration

The "Sensitive" (S) system received better scores than the other two systems in all three parameters (satisfaction, system credibility and frustration), on a scale of 1 (low) to 7 (high): The average of user Satisfaction was 4.75 (compared to system I and M who got 4.42 and 4.46, respectively). system S's Credibility average was 4.83 (compared to system I and M who got 3.92 and 4.54, respectively) and user Frustration average was 3.42 (compared to system I and M who got 3.75 and 4.21, respectively). Participants said that they liked and trusted more the sensitive system since *"the hesitation helped me predict when it is going to err, so I was more prepared"; "The hesitation helps me learn. When the system hesitates, I understand that next time I have to better pronounce the name"*.

The results are presented in Fig. 3.

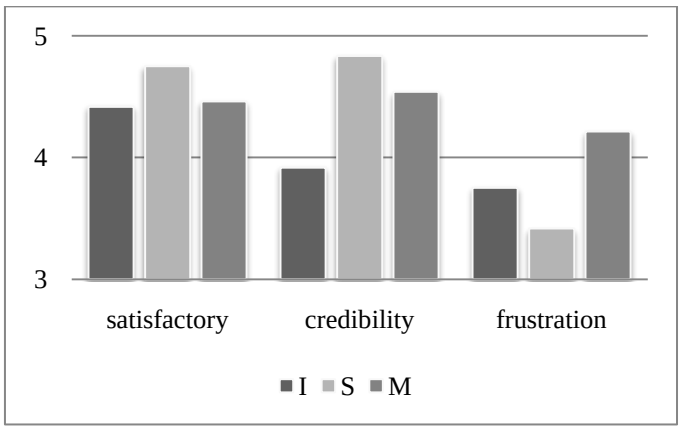

Figure 3. Satisfactory, Credibility and Frustration average in the three systems

C. Comparative preferences

At the final stage of the tests, the subjects were asked to compare the two systems they tested. 64% said that system S was more satisfying than or at least as satisfying as the other system. Fig. 4 demonstrates that when asking to compare between two systems, 32% subjects preferred system S while only 5% subject preferred system M. The equal sign (=) refers to equal preferences.

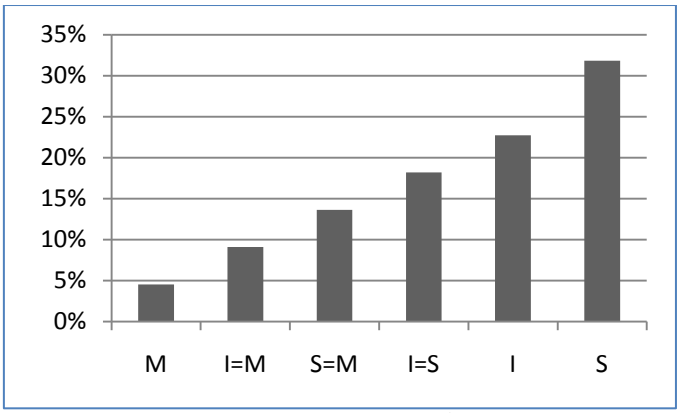

Figure 4. Participants' system preferences.

DISCUSSION

In this paper, we examined whether adding hesitation disfluencies to the system's responds will improve the human-machine interaction. We showed that by adding these prosodic characteristics, the naturalness of the system is improved. Yet, we investigated three different systems that vary in terms of transparency.

The results presented here are preliminary in type and could only manifest a trend. More subjects and maybe a more elaborated experimental setup are needed in order to apply statistical tools and assess significance.

The results suggest that people tend to react more leniently to an "erroneous" answer, i.e., if the answer was given in a hesitating manner, only if the system implies by it that it is about to fail in its recognition. Hence, people prefer *naturalness* as long as it is "real", *transparent*.

The use of verbal hesitation to communicate or express the "cognitive state" of the system was not explored before. Hence, we believe that this work is a first step in improving the human-machine interaction. Exploring this can contribute to other studies, such as [11], where the authors propose to implement human hesitation gestures onto a robot, and to investigate its ability to communicate uncertainty.

We believe it will be interesting to investigate whether different recognition rates can change the user's reaction to the three systems in our experiment. We propose to further test this in future experiments.

REFERENCES

[1] H. R. Pon-Barry, and S. M. Shieber, "Recognizing Uncertainty in Speech," *EURASIP Journal on Advances in Signal Processing*, 2011: 251753.

[2] C. Collins, S. Carpendale, and G. Penn. "Lattice Uncertainty Visualization: Understanding Machine Translation and Speech.recognition," *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization*. Norrköping, Sweden, May 2007.

[3] R. Eklund, Disfluency in Swedish human-human and human-machine travel booking dialogues. Dissertation 882, Linkoping Studies in Science and Technology, 2004.

[4] R. Carlson, K. Gustafson, and E. Strangert, "Prosodic Cues for Hesitation". Dept. of Linguistics & Phonetics Working Papers, 52, 21–24, 2006.

[5] C. Callaway, "Do we need deep generation of disfluent dialogue?" In *AAAI Spring Symposium on Natural Language Generation in Spoken arui Written Dialogue*, Tech. Rep. SS-03-07. Menlo Park, CA: AAAI Press, 2003.

[6] J. R. Benkí, J. Broome, F. Conrad, R. Groves, and F. Kreuter, "Effects of speech rate, pitch, and pausing on survey participation decisions," in *AAPOR meeting*, Phoenix, May 2011.

[7] J. Liscombe, J. Hirschberg, and J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Intespeech 2005*, pp. 1837–1840Lisbon, Portugal, 2005.

[8] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, "Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems," *International Journal of Artificial Intelligence in Education* vol. 16, pp. 171-194, 2006.

[9] M. Corley and R. J. Hartsuiker, "Why Um Helps Auditory Word Recognition: The Temporal Delay Hypothesis," *PLoS ONE* 6(5): e19792. doi:10.1371/journal.pone.0019792, 2011.

[10] E. Aharonson, and V. Aharonson, "Multimodal Interfaces for Mirco User Actions," in *Speech Processing Conference 2011*, Tel-Aviv, Israel, 2011.

[11] A. Moon, B. Panton, H. F. M. Van der Loos, and E. A. Croft, "Using Hesitation Gestures for Safe and Ethical Human-Robot Interaction," *IEEE ICRA 2010 Workshop on Interactive Communication for Autonomous Intelligent Robots*, Anchorage, USA, May 2010.

.