Speech Synthesis and Uncanny Valley*

Jan Romportl

Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Pilsen, Czech Republic
rompi@kky.zcu.cz

Abstract. The paper discusses a hypothesis relating high quality text-to-speech (TTS) synthesis in spoken dialogue systems with the concept of "uncanny valley". It introduces a "Wizard-of-Oz" experiment with 30 volunteers engaged in conversations with two synthetic voices of different naturalness. The results of the experiment are summarized and interpreted, leading to the conclusion that the TTS uncanny valley effect in dialogue systems can probably be superseded and inverted by a positive attitude of the systems' users toward new technologies.

Keywords: text-to-speech synthesis; spoken dialogue system; uncanny valley; experiment

1 Introduction

The concept of *uncanny valley* [1] has been originally introduced by Masahiro Mori in 1970 for description of a typical emotional effect that near-human artifacts elicit in human – for example seeing a very accurate prosthetic hand can provoke feelings of eeriness, whereas seeing an artificial robotic "hand" is as neutral as seeing a real human hand (indeed in its proper functioning conditions as a part of a healthy human body, not e.g. amputated).

Analogically, an archetypical human-style robot, such as Number 5 from the movie Short Circuit or C-3PO from Star Wars, is often perceived almost as cute, whereas a highly accurate near-human robot like Geminoid F, especially when moving, is usually assessed as literally creepy. Mori also notes that the uncanny valley effect is stronger when the artifact is moving.[1] The dimension in which uncanny valley occurs and can be "measured" (at least informally) is called *shinwakan*, a Japanese neologism coined by Mori for what can be translated as "familiarity", or probably better as "affinity" [2], i.e. affinity of a human towards the artifact.

There have been many studies reporting various findings and theories on uncanny valley. Their brief overview is for example in Roger Moore's paper [3] where he also proposes a probabilistic framework for uncanny valley formalization, based on

* This work was supported by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

[1] Due to quite strict space limitation we do not reprint the notoriously known Mori's uncanny valley curve which very well illustrates the effect. However, it is easily accessible e.g. here http://en.wikipedia.org/wiki/Uncanny_valley.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 595–602, 2014.

the presumption that uncanny valley is a specific manifestation of a more general psychological phenomenon called "perceptual magnet effect".

The goal of our paper is to explore (at least to some preliminary extent) links between uncanny valley and text-to-speech (TTS) synthesis used in spoken man-machine dialogue systems. It has been inspired by our previous involvement in the FP6 project Companions where we worked on a highly natural and expressive Embodied Conversational Agent (ECA) for seniors. As a part of this work, we have recorded an audio-visual corpus of 60 hours of spoken dialogues between seniors and a simplified graphical ECA with a rather artificial TTS voice controlled by hidden human operators – so called "Wizard-of-Oz" (WoZ) method [4].

The initial expectations were that the users (seniors) would not truly enjoy the conversations and immerse into them unless ECA is equipped with a highly natural expressive and emotional TTS voice, which was not the case in the aforementioned corpus acquisition. However, to our surprise, the users enjoyed the conversations very much and often got highly emotionally involved in them (no matter their rather artificial conversation partner), as was recently shown by our M.Sc. student Pavlína Heiderová who analyzed in her master's thesis [5] emotional responses of the users of this WoZ-simulated ECA.

As a result of these findings, we posed a question whether it could actually be the *artificial* ("robotic") voice itself that helps the users engage in a *natural* and pleasant conversation with an obviously *artificial* agent. In other words: whether it is possible that by using a highly natural state-of-the-art emotional speech synthesis the system would actually "drag" the users into the uncanny valley, degrading their user experience and conversational comfort by exposing them to a mismatch between different sensory cues (natural voice possessed by otherwise fully artificial entity), distorting their categorization in the aforementioned perceptual magnet effect, just as it is with the contrast between C-3PO and Geminoid F, only this time with speech synthesis. Therefore, we have performed a set of initial experiments aimed at answering this question.

2 Experiments

The experiments were conducted by our M.Sc. student Daniela Tisarová as a part of her master's thesis [6]. The goal of the experiments was to acquire empirical data for supporting or rejecting the aforementioned hypothesis about TTS and uncanny valley.

The experimental protocol was based on "Wizard-of-Oz" simulation of a fictional AI-based small talk dialogue system (chatterbot) having spoken Skype conversation with research volunteers (probands henceforth) over several neutral casual topics, such as public transportation, weather, etc. The dialogue was immediately followed by a structured questionnaire.

The group of the probands consisted of 30 individuals with the average age of 23.5 years, selected mostly from university students of a technical and a philosophical faculty. The group consisted of 15 female and 15 male probands.

2.1 Experimental Protocol

Prior to the experiments, all the probands were briefly introduced to the field of AI-based dialogue systems, their state of the art, problems and challenges. However, they were not explicitly introduced to the field of TTS synthesis and its evaluation, so as to eliminate a potential bias resulting from their knowledge of which particular speech synthesis method is used in the experiment.

The probands were then instructed that they would go through a Skype call with an AI-based spoken dialogue system that will casually chat with them over two photos (a bus station and a railway station) about public transportation. The probands were told that the system has two female identities represented by two different voices and that the Skype call will be divided into two separate conversations, each of them with a different voice and over a different photo. The probands were intentionally given a false idea that the dialogues are for testing purposes of the actual AI-based system performance and that the system is equipped with a state of the art ASR, NLP, dialogue manager and TTS, whereas in fact TTS was the only automatic component really present; the rest was simulated by the experimenter Daniela ("wizard").

The two voices used in the experiment will be henceforth denoted as "Voice A" and "Voice B". Both voices were synthesized by our TTS system ARTIC. Voice A is an old single unit instance TD-PSOLA voice, judged ex post by the probands as being very "robotic". Voice B is a state of the art highly natural unit selection (with no acoustic signal modifications) voice based on a 10+k-sentence corpus. The order of the voices' engagement with the probands was randomized – in 15 cases Voice A was taking part in the first conversation, in 15 cases it was Voice B. The order of the photographs giving a background for the dialogues was randomized along the same lines, too.

The probands did not know that their communication partner is actually the experimenter who was using the TTS system with Voice A/B for synthesizing the turns of the fictional chatterbot and who was mimicking typical behavior of contemporary spoken dialogue systems, such as inappropriate timing of responses (either too long pauses or badly timed barge-ins), lack of common sense (or somehow caricatured common sense), problems with semantical and pragmatical interpretations, etc. – simply, the experimenter mimicked the stereotypes that most contemporary spoken dialogue systems meet and that are usually expected from them by general public. The probands were thus talking to the supposed AI system and were receiving synthesized replies generated by the hidden experimenter.

The experimenter indeed tried to keep the content of all the dialogues across the probands as similar as the individual situations allowed, so that the experiment is not influenced by uncontrolled differences among the dialogues.

2.2 Assessment of Dialogues

Immediately after the Skype call when both conversations ended, the probands were asked to complete a structured online questionnaire. The most important questions were:

1. Which *conversation* was less unpleasant? (with Voice A / with Voice B / no preference)
2. Which *voice* did you like less? (Voice A / Voice B / no preference)

3. Were the conversations interesting for you? (absolutely / mostly yes / mostly no / not at all)
4. How would you assess the Voice A and B, respectively? (robotic or artificial / usual widespread synthetic voice / close to human voice / same as human voice)

Then the questionnaire comprised questions about the probands (age, education, etc.) and also several unstructured questions, such as "How would you assess the dialogue by one word?", "Were you surprised by anything?", "What was the most/least pleasant aspect of the dialogue?", etc. These are, however, beyond the scope of the present paper.

The questions (1) and (2) were formulated in the negative voice because we expected the probands to feel some discomfort or unease in any case (based on our own subjective experience; and it was also an objective of the experiment to build up little tension, otherwise there would be no space for uncanny valley), and so asking them "Which conversation/voice did you like *more*?" could psychologically lead to more frequent frustrated reply "none", meaning simply "I don't like talking to machines at all."

The first two questions show an apparent effort to filter out the probands' opinion on which voice sounds "more/less natural" – we did not ask for naturalness, instead wanted to hear which voice and which conversation caused them more troubles, strangeness or discomfort, which voice in that particular situation "dragged" them more to the uncanny valley. Moreover, we wanted to see if there is a difference in the probands' assessment of their attitude towards the particular *voice* and towards the more complex concept of *conversation*.

3 Results and Their Discussion

The most obvious question that we would like to get answered is indeed whether the probands disliked significantly more Voice A, or Voice B. If Voice B is disliked unequivocally, then there is a clear indication of uncanny valley because the highly natural Voice B does not match properly the typical "machine-like" behavior of the system in all other aspects of the conversation. On the other hand, if Voice A is disliked unequivocally, then the uncanny valley hypothesis for this kind of TTS application can be rejected. However, as Table 1 shows (based on questions (1) and (2)), the results are not unequivocal at all. They are somewhat in favor of Voice B (and rejection of the uncanny valley hypothesis) but they are definitely not convincing. It means that this major question still remains open, as we will discuss further.

The table also shows there is a very accurate complementary relation between "(dis)liking the voice" and "(dis)liking the conversation", which indirectly supports our assumption that the content of all the dialogues is coherent and that the only aspect making the difference is the voice.

We also checked if there is any significant difference between the voice preferences in male and female probands, and by Pearson's chi-squared tests we confirmed the voice preferences are independent on the probands' gender (with $p = 0.74$).

Table 1. Overall preferences of the voices and conversations.

	this voice is less pleasant	conversation with this voice is less unpleasant
Voice A	22	6
Voice B	7	20
no preferences	1	4

3.1 Order of Conversations

Since the preferences (almost one-fourth) in favor of the robotic Voice A must not be neglected, we have investigated more factors that could explain them differently than as an inherent aspect of psychology and cognition of the respective probands.

One of the hypothesis was that the preference could be influenced by the order of the voices that engaged in two conversations with each proband; e.g., if a proband hears the robotic Voice A in the second conversation after being engaged with the natural Voice B for some time, he/she might subjectively feel disappointment purely with the voice qualities, that cannot be cognitively separated from the much more complex attribute of affinity (or Mori's *shinwakan*) towards the voice. However, the chi-squared test shows (again with $p = 0.74$) that we cannot reject the null hypothesis of independence between the voice preference and the order of its respective conversation (the table is thus not necessary here because the values quite closely follow the distribution in Table 1). Therefore, it is quite reasonable to assume that the voice preference is independent on the order of its engagement with the proband.

3.2 Background of Probands

It is quite clear that the background of the probands can influence their affinity towards man-machine conversation – those who are familiar with new technologies, are in active daily contact with them or even are professionally involved in their development are quite likely to react differently when verbally exposed to an AI than those who have either a priori lukewarm attitude towards technology or simply have not been in touch with it.

We have presupposed (among others on the basis of our prior teaching experience at various faculties) that such background differences can very roughly be captured by the field of the proband's study/work – either "rather technical" (including economics), or "rather humanities". Our expectations were quite confirmed by the results of the experiment, as illustrated by Table 2. This contingency table shows the relation between the probands' field of study/work (i.e. field of expertise) and their voice preference. When we stated the null hypothesis as the independence of the voice preference on the field of expertise, we had in this case a reason to reject it by the chi-squared test with calculated $p = 0.05$.

Such a borderline value gives us somewhat medium presumption against the null hypothesis but we must keep in mind that the total number of the probands (here 29 because we have excluded the proband with no preference for/against any of the voices) is still significantly lower than what the rule of thumb says for the Pearson's chi-squared

Table 2. The relation between the probands' field of expertise and their voice preference.

	rather technical	rather humanities	*total*
dislike Voice A	11 (92 %)	10 (59 %)	21
dislike Voice B	1 (8 %)	7 (41 %)	8
total	12	17	29

test (often said to be 50), and therefore the statistical results are not much robust. In any case, it at least points out a very important factor that can be addressed in future experiments.

What is, however, clear is that in this particular experiment the probands with technical expertise quite explicitly disliked conversations with the robotic Voice A. At this moment, we can only speculate about the cause of such an effect – one of the speculations can be that the probands with more technical background easily identified the technical shortcomings of Voice A and that their psychologically default modus operandi can be informally paraphrased as "the more technical shortcomings a thing has, the worse it is". On the other hand, the probands from the field of humanities are more likely to be spared from such technophillic assessments and their preference is driven more by their unconscious affinity towards the communication partner. Such a speculation can thus lead to two (maybe not disjoint) conclusions: 1) people without technical background are more likely to be "dragged into the uncanny valley" of TTS; or 2) people with technical background pay less attention to their unconscious affinity because it is superseded by their technophillic enthusiasm. However, in order to prove or falsify these statements, much more elaborate and extensive experiments are needed.

Since we have found that the voice preference most likely depends on the field of expertise, we wanted to see if the probands' interest in the conversation (question (3) – 15 answers "absolutely", 14 answers "mostly yes", 1 did not answer) could make a difference too. The conclusion is that the null hypothesis "the voice preference is independent on the probands' interest" cannot be rejected by the Pearson's chi-squared test (with $p = 0.40$).

On the other hand, there is perhaps some form of dependence between the probands' field of expertise and their interest in the conversations. We have low presumption against the null hypothesis "the probands' interest is independent on their field of expertise" (with $p = 0.09$), which at least means (given the aforementioned low robustness of the statistics) that this aspect should be further explored in detail. At this moment, we can at least speculate that this again fits well to the image pictured by the previous tests (and also the intuitive stereotypical thinking about the technology users). We do not present the respective tables here due to lack of space.

3.3 Duration of Conversations

Another hypothesis that emerged together with the experiment was a relation between the probands' preference and the duration of their engagement with the voices. What if the probands were getting more frustrated with the robotic Voice A as the conversation took longer? Or the other way around – what if they were getting used to the robotic

Voice A in the course of the conversation, while noticing more and more "uncanny glitches" in Voice B?

Therefore, we have calculated the following quantitative parameters for each proband: a) duration of the whole Skype call, given as mm:ss (average 16:15, standard deviation 03:37); b) duration of the conversation with Voice A (avg. 07:13, stdev. 02:12) and Voice B respectively (avg. 08:02, stdev. 02:39); c) number of turns Voice A (avg. 28.6, stdev. 8.6) and Voice B respectively (avg. 30.7, stdev. 6.5) had in each conversation.

For each parameter and each proband, we have categorized the respective conversation/call into one of three groups: Short, Medium, Long. The Medium category was delimited by the interval of the respective average value minus/plus its standard deviation. The Short category was then indeed everything below this interval and the Long category above. We do not present the tables with the frequencies of each category here mainly due to space limitation and the fact that the most interesting information is given by the aforementioned moments.

We formulated and tested (again by the Pearson's chi-squared test) the following list of null hypotheses:

1. The voice preference is independent on the duration of the Skype call; $p = 0.54$; cannot be rejected.
2. The voice preference is independent on the duration of the Voice A conversation; $p = 0.25$; cannot be rejected.
3. The voice preference is independent on the duration of the Voice B conversation; $p = 0.58$; cannot be rejected.
4. The voice preference is independent on the number of the system's turns in the Voice A conversation; $p = 0.25$; cannot be rejected.
5. The voice preference is independent on the number of the system's turns in the Voice B conversation; $p = 0.98$; cannot be rejected.
6. The Skype call duration is independent on the proband's field of expertise; $p = 0.54$; cannot be rejected.

The conclusion for this point is thus quite clear: the probands' voice preference most likely did not depend in any way on the duration of the conversations.

4 Conclusion

As we have already discussed in the previous section, we have not received unequivocal results. The majority (about three quarters) of the probands preferred the more natural synthetic voice, which speaks against the initial hypothesis of uncanny valley related to "too natural" speech synthesis in spoken dialogue systems. However, still a significant number of probands had the opposite preference, which also cannot be ignored.

Our most important finding here is quite a remarkable influence of the field of expertise of the probands, which leads to our new speculative hypothesis that the "technophillic attitude" of a significant part of the probands covered and superseded their primary *affinity* towards their artificial partner in conversation. We will address this hypothesis in our future experiments that will be aimed at statistically more robust,

balanced and extensive group of probands (especially laypeople outside academic environment).

Moreover, we have shown that the conversation preference is quite well reducible to the voice preference and that the voice preference does not depend on the order of their respective conversations, nor does it depend on the duration of the conversations. The only aspect that made the difference was the background of the probands. This will help future experiments as well.

References

1. Mori, M: The uncanny valley (translated by MacDorman, K.F., Kageki, N.). IEEE Robotics & Automation Magazine, 19(2), 98–100 (1970, 2012)
2. Bartneck, C., Kanda, T., Ishiguro, H., Hagita, N.: Is the uncanny valley an uncanny cliff? In: 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 368–373. Jeju, Korea (2007)
3. Moore, R.K.: A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Nature Scientific Reports, 2(864) (2012)
4. Romportl, J., Zovato, E., Santos, R., Ircing, P., Relaño Gil, J., Danieli, M.: Application of expressive TTS synthesis in an advanced ECA system. In: Proceedings of the ISCA Tutorial and Research Workshop on Speech Synthesis, pp. 120–125. Kyoto, Japan (2010)
5. Heiderová, P.: Perspektivy řečové komunikace mezi člověkem a strojem (Perspectives of Speech Communication Between Human and Machine). Master's thesis, University of West Bohemia, Pilsen (2012)
6. Tisarová, D.: Hypotéza "uncanny valley" ve vztahu k syntetické řeči (The Uncanny Valley Hypothesis in Relation to Synthetic Speech). Master's thesis, University of West Bohemia, Pilsen (2014)