Voice and Trust in Digital Media

Felix Ahlner

Master Thesis
Computational Linguistics
Lund University

June 2008
Supervisor: Agneta Gulz

Abstract

This study investigates to what degree trust affects the results in a choice task. In the task, participants' answers are sometimes manipulated, and we measure how often this is detected. Participants were divided into three groups, each getting the instructions from a different recorded voice which is more or less trustworthy. The results show many kinds of differences between the groups, both in the detection rate and how participants talk about the voices. The conclusion is that trust does affect the outcome of this task, but also that participants can be affected by self-censorship such as political correctness.

Keywords: trust, choice blindness, voice, political correctness

Table of Contents

1 Introduction

The purpose of this study is to investigate how the results of a simple choice task are affected by the participants' trust in the experimenter.

In the original version of this choice task, participants were shown pairs of photographs of female faces, and their task was simply to choose which face they found most attractive. The participants were then shown the face they had just chosen, and at times they were asked to further motivate their choice. On a few occasions, however, the experimenter performed a concealed card trick and switched the two photographs. The participants were thus shown the face they had *not* chosen, but were asked "Why did you choose this face?". The remarkable results are that 3 out of 4 times participants failed to detect this manipulation. This phenomenon is called *choice blindness* (Johansson et al 2005, Johansson et al 2007).

A central hypothesis for explaining the phenomenon of choice blindness concerns participants' *trust*. Firstly, their trust in the invariability of the physical environment is likely to make them unaware of the card trick. Secondly, they are likely to trust the experimenter not to tamper or manipulate the results. Therefore, when participants are asked to motivate the manipulated choices they unconsciously confabulate reasons for choosing a face that they in fact did not choose.

It is well-known that humans readily extract information from voices, such as the speaker's gender, origin, emotional state etc. This process is highly automatic and operates on genuine human voices as well as on recorded voices in computer interfaces (Reeves & Nass 1996, Nass & Brave 2005). The information we extract from a voice helps us to form conceptions about other people, ultimately affecting our trust in the speaker and the message. People generally prefer people that are like themselves – the *in-group* – over people who differ in some relevant aspect – the *out-group*. Language and accent is one such highly relevant aspect (Schneider 2004, Dahlbäck et al 2007).

In order to test the hypothesis about trust, we created a computer version of the choice task in which all instructions were given by a recorded voice – the 'digital experimenter'. 60 participants were divided into three groups, and each of these heard the instructions spoken with a different accent. This allows the digital experimenter to become more or less similar to the participants and therefore more or less trustworthy. After the choice task, the participants are also asked to assess how much trust they had in the experimenter's voice.

Assuming that trust is a factor behind choice blindness we predict that *the number of detected manipulations will differ between the groups*, and that *the explicitly assessed trust will show an inverse correlation with the number of detected manipulations.*

2 Background

2.1 The Media Equation

The most fundamental theory behind this thesis is the equation *Media = Real Life*. This equation, the conclusion of thirty-five studies conducted over several years, was presented in 1996 by Byron Reeves and Clifford Nass in their book *The Media Equation*. The equation means that human beings react to and interact with media in fundamentally the same way as they react to and interact with other human beings. Among other things, studies in the book showed that we respond to reality and filmed reality with the same kind of emotions, and we direct our attention to movement on a screen just as we direct it towards real objects. Furthermore, we treat computer interfaces similarly to how we treat other people: we ascribe personality traits to them and we are polite to them. The social conventions governing our everyday behavior carry over into media – from simple drawn animations and written text up to high-quality photographs and high-fidelity sound recordings (Reeves & Nass 1996).

At first, this idea can seem counterintuitive and absurd, even insulting. As educated adults, familiar with these media, we all *know* that computers and televisions lack emotions and feelings. We *know* that a photograph is just a gathering of pigments on paper. We *know* that the movie we are watching is ultimately a transparent celluloid reel projected onto a screen, accompanied by sound created by vibrating diaphragms. Still, most of us have at some point tried to interact with a computer, printer or television; from a pleading "Come on, it can't take this long …" to "You piece of crap! I swear, if you give me one more paper jam, you're going straight out the window!". Cartoons, books and even photographs can make us laugh or cry, and a good horror movie or thriller affects us in the same way that a tangible physical threat would do.

Some people find it hard to accept that there is such a discrepancy beween our sensible reasoning and our actual reactions and behaviour. So, perhaps this phenomenon applies only to children who are still too young and unexperienced to tell the difference between media and reality? Perhaps it applies to people in the Papuan highlands or the Amazon jungle who are largely unfamiliar with these modern media? Reeves and Nass refute this as well: throughout their studies they found that the Media Equation applies to everyone: children and adults, moviegoers and readers, computer novices and computer experts.

The media equation *can* be overcome, people *can* treat media as a mere representation of the real world – but they tend not to. This type of detachment always requires "a lot of effort […] and [it is] always difficult to sustain. The automatic response is to accept what seems to be real as in fact real." (Reeves & Nass 1996:8).

2.1.1 The method behind the media equation

The studies conducted by Reeves & Nass were all based on earlier studies of human interaction and behavior from the domain of social science: psychology, sociology and communication. Before repeating these studies, one simple yet crucial alteration was made: wherever a study mentioned a principle or conclusion about human behaviour, words like 'person' or 'environment' were substituted with different kinds of media, e.g. "people like to be praised by other people" became "people like to be praised by computers". The study's method and material were then altered in order to make it possible to investigate these conclusions.

It is important to be aware that Reeves & Nass started their research with the intention to find out more about how humans interact with computers, and not to find proof for one specific hypothesis. They readily admit that in the beginning of their research they had no formulated media equation, they rather "believed that people might occasionally confuse media and real life" but that such a phenomenon would be both limited and "curable", i.e. that the more someone has been exposed to media the easier he would have to separate media and reality (Reeves & Nass 1996:6).

2.1.2 The theory behind the media equation

Our human brain is simply "not evolved to twentieth-century technology" (Reeves & Nass 1996:12), it still reacts with the vital 'automatic responses' that have been refined over millions of years in a world without any media. The mammal brain has been evolving over 200 million years, and our social skills as primates go back millions of years. Special human skills such as language go back about 200,000 years.

Given this timeframe, media is something very new. The more advanced 'capturing' media like photography, film and sound recording are no older than 200 years. Througout our entire evolution it has been a simple truth that if something looks, acts or sounds like a human, it *is* a human. This is a very simple rule that was once applicable to our every sensory input.

As mentioned earlier, it takes a lot of conscious effort to set aside these strong ancient automatic responses, and this 'detached' state of mind is hard to sustain. As an example, it is very difficult for someone to distance himself from a scary movie and at the same time keep track of the plot.

Designers of media and interfaces can benefit greatly from these findings. Instead of thinking of media as tools or mere representations of reality, designers can both simplify and improve interaction and interfaces by letting the human social expertise become relevant in media interaction.

2.2 Humans are ‘voice-activated’

By the end of the 1990’s, loudspeakers and even microphones had become integral parts of any home computer, facilitating the spread and development of voice-interfaces. This led Clifford Nass and his new colleague Scott Brave to focus their research on speech, the most social of all our human skills. Humans are experts at analyzing voices to quickly extract ‘social cues’ such as a speaker’s gender, age and emotional state. The speaker’s choice of words and syntactical constructions often provides further social cues, e.g. ‘A mistake was made’ can be perceived as more evasive than ‘I made a mistake’, and ‘Give me a beer’ is ruder than ‘Could I have a beer, please?’.

Again, researchers designed and conducted experiments that were similar to social science experiments concerning human interaction. Subjects ranged from stereotyping of gender and ethnicity to the perception of emotion and personality in a voice. These studies were gathered in the book *Wired for Speech* (Nass & Brave 2005).

As the media equation would predict, the results showed that we perceive and react to talking and listening media as if they were living human beings. The suggested explanation is also similar to what was suggested in the media equation: we have “brains that are wired to equate voices with people” and we “cannot suppress [our] natural responses to speech, regardless of source” (Nass & Brave 2005:3ff). Indeed, when our brain processes genuine human speech it uses the same parts and functions as it does when it processes recorded and even computer-generated speech of lower quality. Nass & Brave summarize this by saying that humans are ‘voice-activated’, meaning that both authentic voices and recorded voices activate our natural responses to human speech.

2.3 Choice blindness

Choice blindness is a phenomenon first described in 2005, and its name is derived from the finding that most people seem to be largely unaware of some of the choices they make, or at least unable to detect when someone manipulates these choices (Johansson et al 2005).

In the original study, the experimenter showed a participant two pictures of female faces and asked him to choose which face they found more attractive. At intervals, the experimenter handed the participant the picture he had just chosen and asked him to verbally motivate his choice. But, in a few of these cases the experimenter performed a card trick to manipulate the participant’s choice, handing him the picture he had *not* chosen. Including all mitigating circumstances (e.g.

that once a participant has detected one card manipulation, he is more likely to detect any ensuing ones) no more than 26% of the card manipulations were detected. In other words, in 3 out of 4 times participants failed to detect the mismatch between choice and outcome, and instead they explained to the experimenter why they had chosen the picture that they in fact did *not* choose.

2.3.1 Trust as a reason behind choice blindness

One suggested reason behind choice blindness concerns our trust in the constancy of our physical surroundings. Things generally remain the same, they do not radically change as in a card trick. One indication of this belief is that in the original study 84% of the participants asserted that they would detect a manipulation of the pictures (Johansson et al 2007). This was of course asked before the experimenter revealed what had really happened. The actual result was that only 60% of the participants detected all manipulations, and that was among the participants that were given unlimited time to look at the two pictures. Among the participants that were given 5 seconds to look at the pictures, *no one* detected all the manipulations. Among those who looked for only 2 seconds, 70% failed to detect *any* manipulation. (Johansson et al 2005). This discrepancy between introspective belief and actual behavior is similar to the participants' reactions in many of the studies in *The Media Equation*.

As with the media equation, one could argue that our trust in the physical world is a result of the environment in which we have evolved. Throughout literally billions of years, inanimate things have tended to stay the same. This simple assumption has been succesful enough, allowing for cognitive resources to be directed to more important and changeable things. Living creatures in all environments take advantage of this assumption through camouflage.

Another kind of trust that is relevant in this case is the trust that participants normally have in the experimenter. Most people will regard him as an authority bound by ethical regulations, and therefore he is likely to be sincere about his intentions with an experiment. This kind of trust can be considered to be culture-specific, e.g. since clothing and other attributes that evoke trust vary greatly between cultures.

2.3.2 Altering of trust

What would happen if the pictures of the female faces were not shown on tangible, physical cards but on a computer screen? Does the trust in the physical world extend to the screen, or would a virtual environment make people more suspicious or watchful, eventually making them detect more of the picture switchings?

This idea was put to the test through a computer version of the decision task, in which a cartoon-like female experimenter showed pictures of female faces (Johansson et al 2007). The results support the hypothesis, since the detection rate in the digital experiment was somewhat higher than in the original experiment: 33% vs. 26%. However, the computer version of the experiment did not use the same set of pictures that were used in the original experiment, so the results from the two tests are unfortunately not statistically comparable.

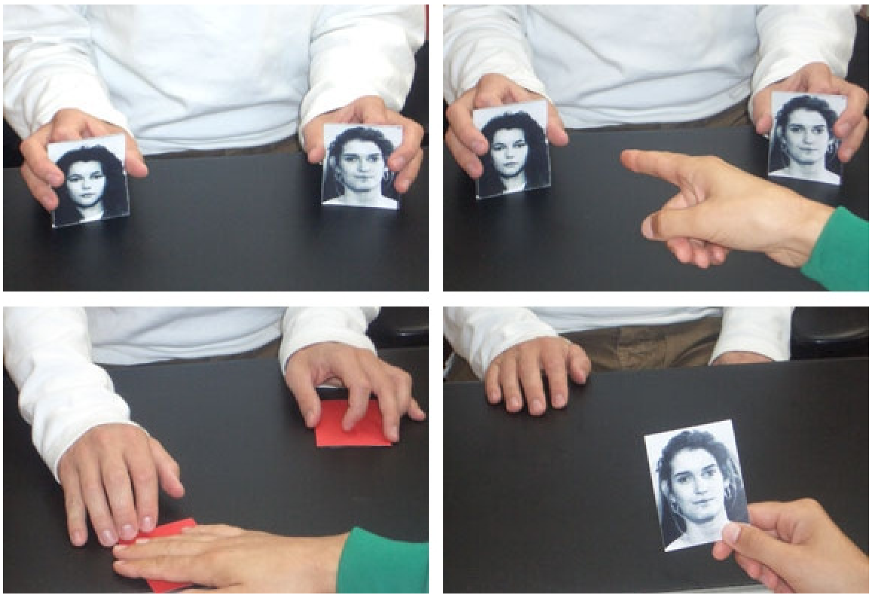

Figure 1. *Layout of the original experiment in Johansson et al 2005*

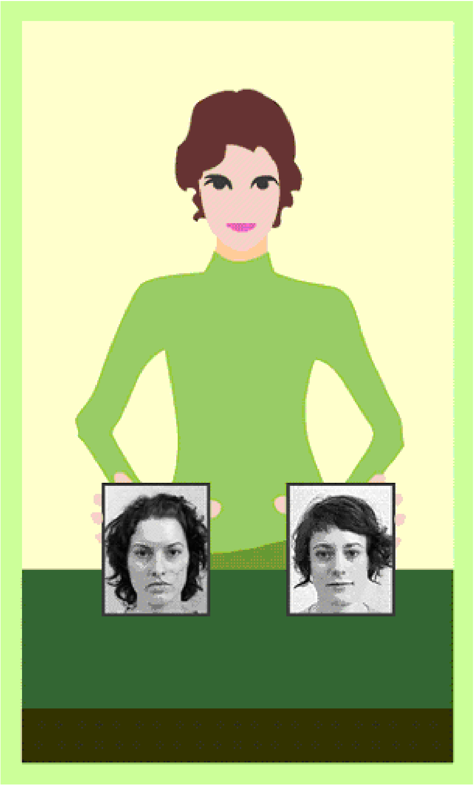

Figure 2. *A screenshot from the digital experiment in Johansson et al 2007*

The participants' trust in the experimenter is also something that can be varied. The experimenter could borrow attributes that normally evoke distrust or watchfulness, e.g. a magician top hat and a wand.

2.3.3 How is trust measured?

In studies of trust one way of measuring is to let a participant perform a task or interact with an interface and then answer gradable questions about his experience, giving a so called *explicit assessment*. This is often followed by an interview or 'debriefing' about the experiment, where the participant can give further comments.

Another kind of measurement is *implicit assessment* where for example participants are given advice on how to perform a task. Their level of trust can be deduced and measured from how faithfully they follow the advice.

Assuming trust to be a factor behind choice blindness implies that the number of detected manipulations is an implicit assessment of the participants' trust in the experimenter and the virtual environment. Manipulated choice tests could then be used as a method for attaining gradable results without having to rely on explicit assessments such as retrospective questionnaires or interviews, which are demonstrably unreliable. Furthermore, as in this study, the numbers of detected manipulations can be compared with the participants' answers in a questionnaire or interview to see if there is any correlation.

2.4 We trust people that are like us

2.4.1 What makes a voice trustworthy?

Many studies have investigated the cross-linguistic acoustic patterns of 'basic emotions' such as fear, anger, joy and sadness. The measured voice parameters usually include intensity, duration, speech rate, pitch variability and pitch mean (Scherer 2003). However, a voice's trustworthiness is a much more complex issue, going beyond mere acoustic patterns.

One of the most basic concepts in trust and distrust is *group belonging*. People trust and favor those who are perceived as belonging to the same group (the *in-group*) but distrust and disfavor those who are perceived as belonging to other groups (the *out-group*). This categorization of people and the subsequent treatment of them is not only an observation that most of us have made in daily life but a well-established topic of research within social psychology and other branches of cognitive science.

The most influential theory about the ‘in-group bias’ is Henri Tajfel’s Social Identity Theory (SIT). This theory argues that our social identity is mainly a product of our group-belongings, and it is therefore in everyone’s best interest to regard his own groups as positive and to act favourable within them in order to strengthen solidarity.

The theory’s indirect explanation of hostility towards out-groups is that one of the easiest, and most common, ways of maintaining a notion of superiority and distinctiveness is to lessen and estrange other groups (Schneider 2004:233f).

Depending on the situation, a person will perceive his membership in certain in-groups as more important than others. Our most basic and involuntary group-belongings such as gender, ethnicity and mother tongue are usually also the most salient and important ones. When the importance of in-groups varies, the importance of the corresponding out-groups changes accordingly.

2.4.2 Groups and trust in voice-interfaces

A famous example of group mismatching within voice interfaces is a talking in-car navigation system from BMW which was launched with a female interface voice. German drivers were “untrusting of a female voice giving directions” and the product was recalled and reworked to better suit the characteristics of the ‘BMW driver’ group (Nass & Brave 2005:55f).

Among the studies in *Wired for Speech*, the one that focused most on trust was a study, similar to this one, about English accents in an e-commerce interface. This interface provided spoken descriptions of four products, all recorded with both an Australian accent and a Korean accent. The recordings also included a few accent-specific phrases like ‘G’day mate’ and ‘Annyong haseyo’.

The participants in this study were divided into two groups consisting of White Americans and first-generation Koreans. Half of the participants in each group were presented with the Australian accent description, and the others were presented with the Korean accent description. After listening, the participants rated the credibility of the descriptions using different adjectives, including ‘trustworthy’. The results showed that the participants put more trust into the descriptions that were given in the accent that most resembled their own (Nass & Brave 2005:61ff).

Another similar study was made with English spoken with American and Swedish accents, and the American and Swedish participants showed the same preference for the in-group accent, even when the Swedish accent gave information about New York and the American accent gave information about Stockholm (Dahlbäck et al 2007). A study with American English and English with a Japanese accent reached similar results (Cargile & Giles 1997).

Dialects or accents also vary in their trustworthiness, and such opinions have obviously developed as a result of historical and social factors that differ between societies and language communities. In terms of how we perceive and regard others, accent and dialect is just as important as gender or visual appearance. This includes aspects such as trust and likeability (Gulz et al 2007).

Since people belong to many different groups simultaneously, a person can be a 'Swedish speaker' in a situation where other languages are present, but as soon as only Swedish speakers are present, he might feel that his identity as a speaker of a particular Swedish dialect becomes more relevant.

3 Material and Method

3.1 Method

The experiment was designed as a between-participant experiment with 3 groups of 20 participants each. All participants performed the same digital choice test, but each one of the three groups were given the spoken instructions by a different voice.

3.2 Material

3.2.1 Choice test

The digital choice test was in the form of a computer application. Instructions for the test were given only in spoken form, uttered by a male voice. The participants could hear the initial instructions over again by clicking a button.

3.2.2 Photos

This version of the choice test used 30 color photographs of female faces. These photographs had not been used in earlier choice tests.

3.2.3 Voices

The three different male voices used in this study were all recordings of the same man, a training actor. The voices were recorded using the same script so they all uttered the same words.

We call the first voice *Standard Swedish* since it spoke Standard Swedish (‘rikssvenska’), an accent deemed neutral and therefore typically used on stage, in newscasts and in instructional messages.

The second voice is called *Fishy Swedish*. It spoke Swedish with a dark, wheezy Stockholm accent. This kind of accent has connotations of alcoholism and criminality.

The third voice is called *Broken Swedish*. It spoke Swedish with a mix of Polish and Russian accents. This accent manifested itself mostly in intonation but also in individual sounds (such as replacing [h] with a harsher [x] or pronouncing long a’s [ɑː] as [aː]). This voice was recorded with the same script as the two preceding voices, so it used the same words and uttered no ungrammatical constructions.

Earlier digital experimenters have sometimes been very repetitive, something that can strain the participant’s concentration and interest. In this flash application no soundclip was ever repeated, but instead each image pair had its own designated soundclip with the occasional comment like “now we have gone through half of the image pairs” or “here comes the next to last image pair”.

3.2.4 Questionnaire

This questionnaire was based on the questionnaire that was embedded in the computer application used in Johansson et al 2007. In this study it was used in a face-to-face interview and extended with questions about the voice.

3.3 Participants

The 60 participants (31 female) were all native speakers of Swedish and most of them came from southern Sweden (from or south of Stockholm). This makes them members of the 'speakers of Swedish' group. The average age was 23.5 with a median age of 22.5.

3.4 Procedure

3.4.1 Pretence

In order to make participants disregard the experimenter's voice, they were told that the experiment was part of a study about the attractiveness of female faces.

3.4.2 Choice test

After the participants had arrived they were told that the computer would give them their instructions and were then left alone in the room with the computer to perform the choice test. A randomization script decided which voice the participant heard.

After hearing the initial instructions, the participants were presented with a total of 15 image pairs, each being shown for 5 seconds. On 6 occasions they were asked to further motivate their choice by indicating the importance (much, a little, not at all, not sure) of certain facial features (face, eyes, hair, smile). There was also a text field for additional written comments. 3 of these 6 image pairs were manipulated (trials 7, 10 and 14).

After the participants had seen, chosen and motivated the pictures, the digital experimenter's voice thanked them for their participation and said good bye.

3.4.3 Questionnaire

The participants then joined the experimenter in an adjacent room for a short interview with both general questions and gradable questions.

The first question was a very general question about the experiment and about how interesting the participants had found it. After that came questions about the voice, first the more general "What did you think about the experimenter's voice?" and then "Did you have confidence in the voice?". This question was gradable from 1 (weak confidence) to 4 (strong confidence).

Thereafter came two questions which allowed the participant to reveal any detection of pictures being switched. The first question was the general question "Did you feel that anything was strange about this experiment?". If the participants had not shown any sign of detection at this point, they were asked the tongue-in-cheek question “Next week we will perform a study similar to this one, but we will secretly switch the pictures so that people are asked to motivate the picture they *didn’t* choose. Do you think that you would detect such a change?”. The participants were then asked to grade how likely it was that they would detect this, from 1 (I don’t think I would) to 4 (I’m certain I would).

Finally came the disclosure that the pictures had been switched already this time, and the question “Did you detect it?”. If the answer was positive, the final gradable question was “How many times did you detect it?”.

Directly after this, the experimenter assured the participants that it was rare for people to detect the switch. The final questions concerned the participants’ mother tongue and whether they had seen or heard about this experiment before. Finally, the participants signed a form of consent.

The reason for doing the interview with a human experimenter in another room goes back to one of the studies in *The Media Equation*, namely that humans are polite to computers just as they are to other people (Reeves & Nass 1996:19ff). In order to make our participants feel free to evaluate the experiment and the voice with full honesty, we distanced them first from the computer experimenter through its saying “good bye”, and secondly from the computer by doing the interview in another room.

One advantage of having a spoken, interview-like questionnaire over a paper questionnaire is that the latter gives the experimenter no information about the participants’ thoughts during the answering. If the participant has been completely oblivious about the picture switching, he might become embarrased once he reaches the final question. He might even go back and change his earlier answers e.g. under the question “Did you feel that anything was strange about this experiment?”. In an interview, on the other hand, the experimenter can observe the participant’s reactions and often notice signs of doubt or certainty in his voice.

3.5 Levels of detection

The layout of the experiment gave the participants several opportunities to show that they had detected any pictures being switched. As in earlier versions of the choice task, these opportunities are divided into ‘levels of detection’, with the first level being the most important.

The first level is called ‘concurrent detection’ and applies to cases where participants immediately report that the pictures had been switched or that ‘something went wrong’. Participants are left alone during the experiment and can only report in writing, therefore this level also includes comments that participants give directly after the experiment is over, or as an answer to the first question “What did you think of the experiment?”.

The second level is called ‘retrospective detection’ and it applies to cases where participants reveal signs of detection in their answers to the question “Did you feel that anything was strange about this experiment?” or to “Do you think that you would detect such a change?”.

The third level is called ‘possible retrospective detection’ and applies to cases where participants do not reveal any sign of detection until after the experimenter has revealed that some pictures were switched. In these cases, the participants are asked to describe the faces they claim have been manipulated.

4 Results

The results show a significant difference in detection rate between Standard Swedish and the other two groups, something that supports our initial theory. However, the explicit assessments do not correlate very well with the detection rate.

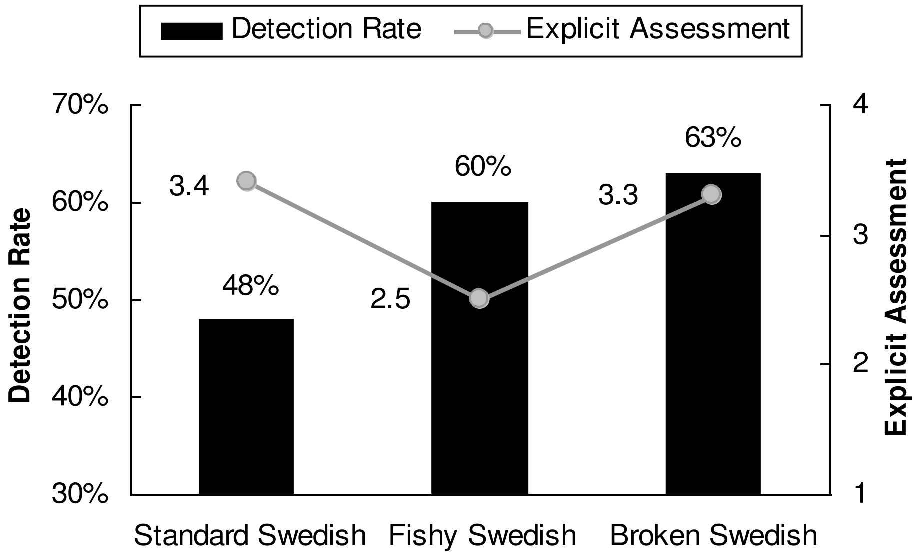

Figure 3. The results from the study

The detection rate for the Standard Swedish voice was 48%, a high number compared with earlier studies (see 5.2). As hypothesized, the detection rates for the other two voices were even higher: 60% for Fishy Swedish and 63% for Broken Swedish.

The explicit assessment also divides the three voices into two groups, but not along the same lines. The assessment, rated on a scale from 1 to 4, reaches only 2.5 for Fishy Swedish, but the numbers for Standard Swedish and Broken Swedish show no significant difference, rated as 3.4 and 3.3 respectively.

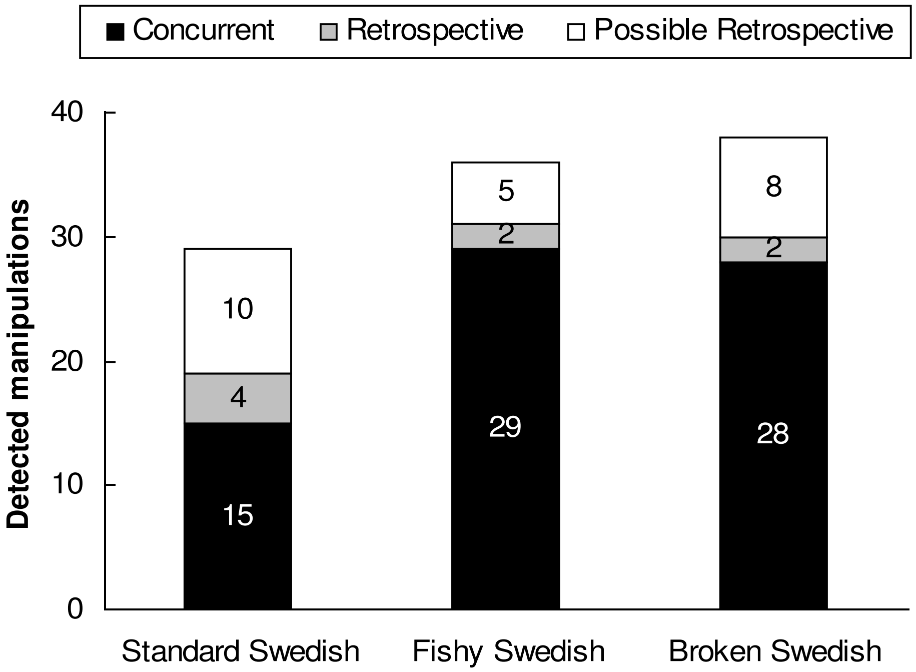

Figure 4. Detections divided into levels (see 3.5)

When the detections are divided into their three ‘levels of detection’ (as defined in 3.5), another distinctive pattern emerges. Not only do Fishy Swedish and Broken Swedish have a higher detection rate, they also have a larger proportion of concurrent detections than Standard Swedish. A high number of concurrent detections indicates that participants are watchful and suspicious.

The concurrent detections make up only 52% of the total number of detections for Standard Swedish, whereas they make up 81% of the detections for Fishy Swedish and 74% for Broken Swedish.

5 Discussion

The results support the initial hypothesis in almost all aspects. The only deviation is the case of Broken Swedish, for which the hypothesis predicts a lower explicit assessment.

5.1 Reactions to the voices

Standard Swedish and Broken Swedish received almost identical assessments, but the participants' comments about the two voices differed in nature.

Standard Swedish was often described as "nice", "good" and "pleasant". Other common descriptions centered on its neutrality, and it was often called "the typical voice used for instructions". Some participants had to think for a while when answering questions about this voice, simply because they had not reflected about it at all.

Broken Swedish was also occasionally called "nice", "pleasant" and "distinct". However, many participants said that the voice was "atypical for this context" and that they immediately had become aware of the accent.

Fishy Swedish was perceived as part fishy and part comical; most participants smiled or laughed when they were asked about the voice. Some participants said that it sounded like a hobo or a drug dealer, but just as many called it "entertaining" or "funny" and likened it to various Swedish comedians and actors. What these two views have in common is that they both perceive the Fishy Swedish voice as unserious and extremely atypical for this context.

5.1.1 Does political correctness affect the assessment?

One theory behind the unexpectedly high assessment of Broken Swedish is that participants have censored themselves in order not to express opinions that others might find offensive, i.e. *political correctness*. The *Concise Oxford English Dictionary* defines political correctness as "the careful avoidance of forms of expression or action that are perceived to exclude or insult groups of people who are socially disadvantaged or discriminated against" (Soanes & Stevenson 2004).

The assessment ratings also go hand in hand with the comments that participants gave: Fishy Swedish received several explicitly distrustful comments whereas Broken Swedish only received one such comment. Some participants also asserted that they had "no problem with immigrant-Swedish", even though they had never said anything to the contrary. Participants would give such comments both before and after they had learned the purpose of the study.

5.1.2 What is it that we distrust?

Another theory concerns the detection rates rather than the assessments. If we assume that the participants' assessment are not affected by political correctness, then perhaps the detection rate does not reflect their trust in the *voices*, but in something else?

Both Broken Swedish and Fishy Swedish were described as “atypical” or “non-standard” voices for giving formal instructions, and participants who heard these voices often said that they reacted to this immediately. These voices in this particular context might have made participants suspicious of the *situation* rather than suspicious of the individual voices. A suspicion towards the entire situation could also be an explanation to why concurrent detections were so common for both Broken Swedish and Fishy Swedish.

This theory would explain the actual results of this study, but at the same time it conflicts with the many other studies showing that people trust people who sound like themselves more than they trust others. Why is Fishy Swedish, spoken by a native albeit atypical voice, assessed much lower than Broken Swedish, and why is Broken Swedish rated as equal to Standard Swedish? When this rather crucial conflict with earlier studies is taken into consideration, the theory of political correctness seems more plausible. That theory would emphasize distrust in particular voices as the explanation of the high levels of concurrent detections.

5.2 Ability to compare with earlier studies

The comparisons within this study posed no problem, but it eventually became clear that this study could not be directly compared with earlier studies since some of the variables in the studies’ layout are incompatible.

To begin with, the photographs used in this study were shown in color, while earlier studies had used grayscale photographs. Color adds more information to the photographs, making it harder to confuse them with each other. Most of the photographs also showed a little piece of clothing, so even if the faces are very similar, the color of the clothes may differ in ways that would not have been discernible in grayscale.

Another thing to take into consideration is that the photographs used in this study were shown bigger than the ones used in earlier studies. These two variables combined form one possible explanation to why the detection rates were so high in this experiment. Another possible contributing factor, as suggested in the introduction, is that people trust a virtual environment less than they trust a tangible physical environment.

5.3 Conclusion

The results from this study strongly support the theory that trust is a factor behind choice blindness, and they also corroborate the many earlier findings that show that we trust voices that are similar to our own more than we trust other voices.

The results also suggest that people's assessment of trust can be affected by factors such as political correctness. This tendency stresses the need for a tool that can obtain a measurable implicit assessment of trust, and choice blindness seems to be such a tool.

The dimension of trust, and how it can be measured, should be further investigated in between-group studies. There are many variables that can be varied: gender, ethnicity, accent, and even clothing.

Once again it becomes obvious that software designers who develop voice interfaces should always investigate what kind of voice fits best in a specific situation. Especially when an application needs to establish a bond of trust with its users, a wisely chosen voice is a crucial component.

Acknowledgments

I want to thank my supervisor Agneta Gulz for her ideas and her support throughout this project. I would also like to thank Magnus Haake for creating the application used in this test, and Nils Derndevik for creating and recording the three voices.

6 References

Cargile & Giles 1997 Cargile, A.C. & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language and Communication 17:3*, 195–217.

Dahlbäck et al 2007 Dahlbäck, Nils; QianYing Wang; Clifford Nass; Jenny Alwin (2007). *Similarity is more important than expertise: accent effects in speech interfaces*. In: *Proceedings of ACM CHI 2007 Conference on Human Factors in Computing Systems*, pp. 1553–1556.

Gulz et al 2007 Agneta Gulz, Felix Ahlner, Magnus Haake (2007). Visual Femininity and Masculinity in Synthetic Characters and Patterns of Affect. Affective Computing and Intelligent Interaction 2007:654–665

Johansson et al 2005 Johansson, Petter; Lars Hall, Sverker Sikström, Andreas Olsson (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science* 310:5745, pp 116–119.

Johansson et al 2007 Johansson, Petter; Lars Hall, Agneta Gulz, Magnus Haake, Katsumi Watanabe (2007). Choice blindness and trust in the virtual world.

Nass & Brave 2005 Nass, Clifford & Brave, Scott (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge: The MIT Press.

Reeves & Nass 1996 Reeves, Byron & Nass, Clifford (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York.

Scherer 2003 Scherer, Klaus R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40:1–2, pp 227–256.

Schneider 2004 Schneider, David J. (2004). *The Psychology of Stereotyping*. NY: Guilford press

Soanes & Stevenson 2004 Soanes, C. and Stevenson, A. (2004), *Concise Oxford English Dictionary*, Oxford University Press, Oxford.