Speech Melody Matters – How Robots Profit from Using Charismatic Speech

Kerstin Fischer[1], Oliver Niebuhr[2], Lars C. Jensen[1] and Leon Bodenhagen[3]
University of Southern Denmark

Corresponding author: Kerstin Fischer

University of Southern Denmark
Department of Design and Communication
6400 Sonderborg
Denmark
kerstin@sdu.dk
+45-6550-1220

Abstract

In this paper, we address to what extent the proverb 'the sound makes the music' also applies to human-robot interaction, and whether robots could profit from using speech characteristics similar to those used by charismatic speakers like Steve Jobs. In three empirical studies, we investigate the effects of using Steve Jobs' and Mark Zuckerberg's speech characteristics during the generation of robot speech on the robot's persuasiveness and its impressionistic evaluation. The three studies address different human-robot interaction situations, which range from online questionnaires to real-time interactions with a large service robot, yet all involve both behavioral measures and users' assessments. The results clearly show that robots can profit from using charismatic speech.

Keywords:

human-robot interaction; prosody; social functions of speech; charisma; robot personality; persuasion

Introduction

Communicating with robots via speech is becoming increasingly important, to reach populations with special needs (e.g. people with dementia, cf. Rudzicz et al. 2015), to allow seamless interaction when people's hands are not free, or simply because people prefer interaction via speech (Grigore et al. 2016). Speech interfaces involve both speech recognition and speech synthesis – that is, the robot is both understanding and producing speech. In this paper, we focus on robots' speech productions, in particular the fact that saying something always involves saying it in one way or another. For this reason, speech melody is necessarily present in robot speech as much as in human speech.

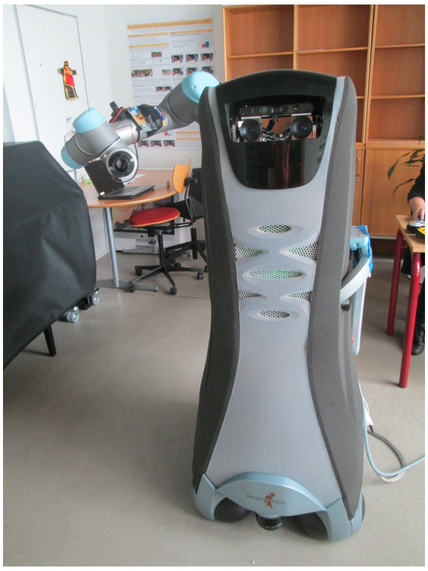

Figure 1: Care-O-bot

In robot speech, usually off-the-shelf speech synthesizers are used (e.g. Mary TTS, Festival, eSpeak, FreeTTS, Vocoder, HTS, etc.), which come with speech melodies that depend on the sentence structure, for instance, by placing stress on content words (cf. Taylor 2009; Dutoit 2013). These speech synthesizers are not implemented with persuasiveness in mind, but target as many possible uses as possible: "commercially developed TTS system have emphasized coverage rather than linguistic sophistication, by concentrating their efforts on text analysis strategies" (cf. Dutoit 2013:148). Thus, speech melody is not adapted to different purposes, and most systems do not even allow the manipulation of prosodic features, such as speech melody, duration, spectral distribution, formant dynamics, etc. Some studies in HRI have therefore used human recorded speech (e.g. Nishio et al. 2012; Kanda et al. 2008), and several studies show that people prefer human speech over presynthesized speech (e.g. Nass & Brave 2005). Nevertheless, in the long run, with robots becoming more flexible and situations of use becoming more varied and unpredictable, robots will have to synthesize their own speech and cannot rely on pre-recorded utterances alone.

In this paper, we present three studies that investigate the influence of speech melody and other prosodic features in robot speech both on users' impressionist evaluations of the respective robots and on the robots' persuasiveness. The results show that speech melody, and prosody in general, has a considerable effect on how robots are evaluated and how persuasive they may be; this finding proved to be robust in that it holds across HRI scenarios and robots. These results indicate that robot designers should pay careful attention to how robots say what they have to say if we want robots to function seamlessly in social interactions.

Previous Work

Previous work indicates that various characteristics of robot speech may play a role in human-computer and human-robot interaction (cf. Nass & Brave 2005). Especially whether the robot's voice is human or whether it is synthesized seems to make a considerable difference. For instance, Stern et al. (2006) compare attitudes to, and the persuasiveness of, human and computer synthesized voices, and while they do not find any influence on persuasiveness, their results show a clear preference for human voice. Similarly, Walters et al. (2008) compare participants' preference for spatial distance to a mechanically looking robot depending on the robot's voice; they have the robot either speak with a human, recorded, male voice, with a human, recorded, female voice or with a synthesized voice. They find consistently that people prefer the robot to be furthest away when it uses a synthesized voice; in contrast, when the robot uses human voices, they come even closer than when they approach a person.

A possible explanation for this finding is provided by an analysis by In & Han (2015), who investigate the use of synthesized speech for robot-assisted language learning. They demonstrate that the F0 range of utterances, i.e. the range in speech melody between between the lowest and the highest points, produced by a common text-to-speech system is much narrower than by a native speaker and thus does not serve as useful learning stimuli for language learning – in fact, learners even corrected their own speech in the wrong direction. In contrast, students learned effectively to increase their speaking rate from repeating after the robot if it used a native-like speech rate. Nevertheless, the robot did not produce adequate adjustments of the speech rate depending on the context. The authors conclude that currently robots do not function as adequate language learning tools if their speech characteristics do not match native-like characteristics in speech melody and variability.

Nass & Brave (2005) provide the broadest analysis of the effects of robot speech characteristics on human-robot interactions. They analyze aspects of prosody, such as gendered speech, but also the use of particular linguistic features, such as the use of 'I' by artificial systems. In general, they find for both recorded and synthesized speech that the more feminine-sounding the voice is, the more female gender stereotypes will be attributed to, and expected of, the artificial speaker, and conversely, the more masculine-sounding the voice is, the more male stereotypes it will evoke. They also find interaction effects between users' personalities and properties of synthesized voices; similarity in gender and emotional state in the expression of introversion and extroversion in voice interfaces leads to greater liking and compliance; in contrast, similarity in the voice qualities themselves does not (2005: 39). In any case, Nass & Brave find voices to consistently convey personality, synthetic as much as natural ones. Furthermore, with respect to emotion expression, they find that contents of stories are liked better if they are presented by a voice that matches the content emotionally (Nass & Brave 2005: 88). Thus, a funny story is funnier if it is presented by a happy voice. Since voices are never truly emotionally neutral, emotional expression in voice interfaces needs to be designed purposefully (2005: 95; see also Jung 2017).

Finally, Nass & Brave (2005: 133) find that interfaces with inconsistent cues, such as a human face matched with a synthesized voice and a synthetic face with a human voice, are consistently evaluated as stranger, more rude and more upsetting (see also Powers & Kiessler 2006). Similarly, regarding the formality of the robot's speaking style, Goetz, Kiessler & Powers (2003) investigate the relationship between formal and informal, playful conversational styles and the areas of applications of the respective robots and find mismatches to receive lower ratings.

Crumpton & Bethel (2016) review the state of the art with respect to emotional expression in robot speech. They argue that there is as yet "very little research into how manipulating a robot's voice would affect its users" (2016, p. 280). One such study is Leyzberg et al. (2011), which finds robots that express emotions through (pre-recorded) speech to elicit better teaching input from the users. Furthermore, Tielmann et al. (2014) find children to respond more emotionally themselves if the robot expresses emotion verbally. Jung (2017) argues that robots need to align their emotional expression with users in order to participate in socially acceptable interactions with humans.

Similarly, in a series of studies involving non-linguistic utterances, Read and Belpaeme (e.g. 2010, 2014) show that prosodic-expressive speech characteristics matter, such that people interpret them categorically with respect to certain emotional states, yet that these interpretations largely rely on the context, for instance the activity the robot is involved in.

There are also some studies that address robot speaking styles more broadly; for instance, Andrist et al. (2013) develop a framework to investigate the effects of particular linguistic features on robot persuasiveness directly. They design a scenario in which two robots are placed on each side of a monitor on which participants see pictures of possible sites to see, and the two robots make suggestions as to which site the participant should choose for a fictional visit on a tour. The behavioral measure is then the extent to which participants follow the robots' advice. Order of presentation of the stimuli was counterbalanced during the experiments so that participants heard each manipulation from both robots equally often. The language features investigated in that paper concern linguistic cues to expertise; in particular, the suggestions for sites were manipulated regarding the robot signaling that it wants the best for the listener and that it has experience from prior helping, as well as regarding the logical progression of the arguments presented, the accessibility of the information through metaphor use and the fluency of the robots' speech. The results show clearly that participants follow those robot's suggestions more that exhibits such cues to expertise. Another speech phenomenon investigated in this way is dialect (Andrist et al. 2015); in these experiments, the authors analyze the interaction between the effects of local dialect versus standard language and cues to expertise. They find expertise cues to be persuasive across cultures, but also local dialect and standard language to be persuasive to different degrees depending on the kinds of expertise cues used.

However, there are also studies that find no effects for speech characteristics, in particular, for speech melody; for instance, Mutlu (2011) manipulates prosodic characteristics of synthesized utterances in order to make the robot sound more lively and expressive; contrary to expectation, the robot that uses the more expressive speech style is not more persuasive than the one with the unmanipulated speech style, in contrast to all other verbal and nonverbal cues investigated, which all have an influence on the robot's persuasiveness. Likewise, Chidabaram et al. (2012) report a study which is similar to the studies reported on here in that they also investigate the effects of prosodic differences in speech style on the persuasiveness of a robot, and they also use synthesized speech which they manipulate in order to make it more expressive. In particular, they manipulated the vocal tone parameters of the synthesized speech, creating two versions of each utterance that are identical in verbal content but differ in intonation so that one was highly monotonous and the other was highly expressive. A manipulation check shows that listeners rate the more expressive speech as marginally more expressive. The authors then study the effects of these manipulations in pitch, either alone or in combination with nonverbal behavior of the robot. The behavioral measure used is whether participants state that they intend to follow the robot's advice. The authors do however not find any effect for the different speech styles. One possible explanation for these null results is that the researchers in both studies manipulated only pitch (or its main acoustic correlate F0); human persuasive (or charismatic) speakers differ in a complex bundle of melodic features, not just intonation (Niebuhr et al. 2016ab), and perception experiments based on listener ratings of manipulated and resynthesized speech stimuli showed that the impression of persuasiveness relies on multiple parameters, with pitch not even being the primary one (cf. Landgraf 2014; Berger et al. 2017). Taking this fact into account, the present study also created multi-parametric differences in the robots' speech stimuli.

To sum up, there is surprisingly little research on the effects of prosodic characteristics of robot voices, and some studies are rather discouraging, pointing to various problems with synthesized speech. At the same time, rather few linguistic dimensions of speaking styles have been investigated (e.g. human vs. synthesized voice, lively vs. monotonous; female vs. male; formal vs. informal). In human interaction, speech styles serve many additional functions, such as engagement, responsivity, social hierarchy, emotional expression, the expression of certainty and of personality, to name but a few (Stibbard 2001; Campbell & Mokhtari 2003; Schuller et al. 2015; Niebuhr 2017). In the current study, we investigate the effects of manipulating robot speech melody and other prosodic features on robot persuasiveness. Furthermore, we address how the way the robot is perceived by participants is influenced by different speech melodies.

Experimental Studies

Different versions of robot speech were created by synthesizing the robots' speech using Mary TTS, a standard, public domain speech synthesizer, and then by manipulating the speech melodies to correspond either to the speech characteristics of Steve Jobs (SJ) or to the speech characteristics of Mark Zuckerberg (MZ). This manipulation relies on the linguistic descriptions of the prosodic patterns of these two famous and successful CEOs (cf. Niebuhr et al. 2016b). We created two variants of the same synthesized speech file by adjusting characteristic aspects of the speech by Steve Jobs and Mark Zuckerberg using the speech analysis software *praat* (Boersma 2001). We then investigated the effects of this manipulation in three different situations, all involving both behavioral and impressionistic evaluation.

We carried out three different studies in three different scenarios since there is reason to believe that robot behavior may be judged differently depending on its presentation. For instance, real-life interactions may differ from interactions with online characters (cf., for instance, Bainbridge et al. 2012), and first person interactions may differ from third person presentations of stimuli, i.e. situations in which speakers observe human-robot interactions (see Strait et al. 2014). In order to show that the effect of robot speech melody is really pervasive, we illustrate it here on as different scenarios and robots as possible. Accordingly, in order to investigate persuasiveness, we use several different behavioral measures.

Scenario Choice

The first study was carried out as an online questionnaire study in which the robot addresses the participant directly. The robot used is a small, relatively abstract humanoid robot, JD's EZ-bot (see Fig. 4). The behavioral measure is whether participants follow the robot's request to take the long questionnaire (rather than a shorter version). The second study uses the same scenario as Andrist et al. (2013, 2015); that is, two robots make suggestions for a possible trip to Paris, between which the respective participant has to choose. The robots used in this study are two Keepon robots (see Fig. 6). The third study involves the large service robot Care-O-bot (Graf et al. 2009, see Fig. 1), which lectures participants about the effects of certain life-style decisions, in particular: performing regular health checks, taking the stairs and eating less sugar. The behavioral measures are accordingly whether participants allow the robot to check their blood pressure, whether they take the stairs or the elevator to the next experiment, and whether they decide for fruit or sweets offered to them. To sum up, the experiments were designed to cover a broad range of different human-robot situations, different robots and different behavioral choices in order to demonstrate the effects of speech characteristics across scenarios. Furthermore, in all studies participants' impressionistic evaluations of the robots were elicited.

Models of Speech Melody

In order to manipulate the robots' voices in terms of their melodic characteristics, we had to find points of reference in the acoustic parameter space that are, firstly, clearly distinct from one another (i.e. associated with changes that exceed the known Just Noticeable Differences (JNDs) in speech perception and that can, secondly, be related on external empirical grounds to persuasiveness and correlated personality traits. JNDs, also known as difference limens, represent limitations of the human ear in the detection of variation among an acoustic parameter. Much experimental research in the field of speech prosody has been dedicated to determining these JNDs (see Moore 2013 and Niebuhr et al. 2019 for summaries), and inter-individual differences, such as the one between musicians and non-musicians, suggest that JNDs not only reflect physiological or biological properties but are at least partly learned, i.e. experience-based thresholds (Gregory 1997). With respect to pitch, for example, pitch differences need to exceed two semitones to be reliably perceived by listeners. Pitch slopes need to differ by at least 60 % to sound more/less steep for listeners, and too shallow slopes (in terms of semitones per second) appear to be flat (Niebuhr et al. 2019).

We decided to use the analyses of the speech of two famous CEOs as the two points of reference in our study. One of them concerns the melodic characteristics of Steve Jobs (former CEO of Apple) who is, across many independent scientific and economic sources, considered "a master of the art of effective and persuasive speaking" (http://www.wbs.ac.uk/news/why-steve-jobs-was-such-a-charismatic-leader1/, see also Heracleous & Klaering 2014). The other point of reference is constituted by the melodic characteristics of Mark Zuckerberg (current CEO of Facebook). Both Jobs and Zuckerberg are powerful and influential men, and thus their ways of speaking should generally be acceptable for listeners in speech-perception experiments. Nevertheless, at a more detailed level, Zuckerberg may be viewed as the prosodic antithesis of Jobs, at least in terms of his speech communication or presentation performance.

Recent studies by Niebuhr et al. (2016a,b) indicate that Steve Jobs and Mark Zuckerberg constitute the opposite poles of those prosodic dimensions that characterize charismatic speech, with other public speakers, like Oprah Winfrey, being localized between these poles. The authors show that the prosodic features of keynote speeches of the two famous CEOs are associated with significantly different melodic configurations comprising a large range of acoustic parameters. The melodic features involved are similar to those that are known from analyses of advertisements and politicians' speeches to be related to persuasiveness and positive character traits like enthusiasm, passion, charm, and convincingness (Gelinas-Chebat et al. 1996; Rosenberg & Hirschberg 2005, 2009; Biadsy et al. 2007; Nienhuis 2009;

Pejčić 2014). The acoustic-melodic profiles of Steve Jobs and Mark Zuckerberg that were worked out by Niebuhr et al. (2006) provide the basis for the creation of our stimuli.

Stimulus Creation

Raw speech stimuli were generated using Mary TTS (Schröder & Trouvain 2003), a standard, public domain speech synthesizer. These stimuli were then further manipulated by means of PSOLA pitch and duration changes in *praat* (Boersma 2001), and subsequent acoustic-energy changes and hesitation insertions were made using Audacity (www.audacityteam.org/). The manipulation aimed at changing the raw stimuli towards the acoustic-melodic profiles that were determined for Steve Jobs and Mark Zuckerberg in the studies by Niebuhr et al. (2016a,b). That is, two versions were created for each automatically synthesized robot utterance, one resembling the acoustic-melodic profile of Steve Jobs, and the other resembling the acoustic-melodic profile of Mark Zuckerberg.

The manipulation procedure involved the following acoustic-melodic parameters: Pitch level (measured in semitones, st), pitch range (measured in semitones, st), acoustic-energy level (RMS, measured in decibel, dB), speaking rate (measured in syllables per second, syl/s), emphatic accent frequency (measured in counts per minute, cpm), hesitation frequency (measured in counts per minute, cpm), silent pause duration (measured in deciseconds, ds), and pitch-accent preference (measured in counts per minute, cpm). Note that this is only a subset of all parameters that characterize the different speaking styles of Jobs and Zuckerberg and which are potentially relevant for their deviating perception in terms of persuasiveness and positive character traits. The manipulated subset was selected according to three criteria: (1) The first criterion is feasibility; for instance, differences in voice-quality are very likely to have a strong impact on perceived persuasiveness and positive character traits (cf., for instance, Signorello & Demolin 2013 and Niebuhr et al. 2018). However, voice-quality parameters are hard to implement and manipulate in a consistent and natural-sounding way in stimuli with current speech-signal processing (i.e. resynthesis) tools. Voice quality was therefore not manipulated in the current data. (2) The second criterion is prototypicality; in particular, pitch, pitch timing, acoustic energy (i.e. loudness), tempo, and phrasing or pausing are unquestionably the core features of speech melody that should get priority over other parameters (see Ladd 2014 for a discussion about the defining parameters of prosody). (3) The third factor is control; that is, parameters like prosodic-phrase duration, rhythm, and pause frequency cannot be varied without simultaneously changing the wording or syntagmatic structure of the stimuli. This would have introduced potentially confounding variables into our experiment that could undermine its internal validity.

The parameters chosen were shown to be key factors for a speaker's charismatic impact on listeners in our previous work, in particular, F0 range, tempo, and emphatic-accent frequency (Niebuhr et al. 2017); Berger et al. (2017) have conducted parameter-specific tests that demonstrate the influence of each of these factors on charisma.

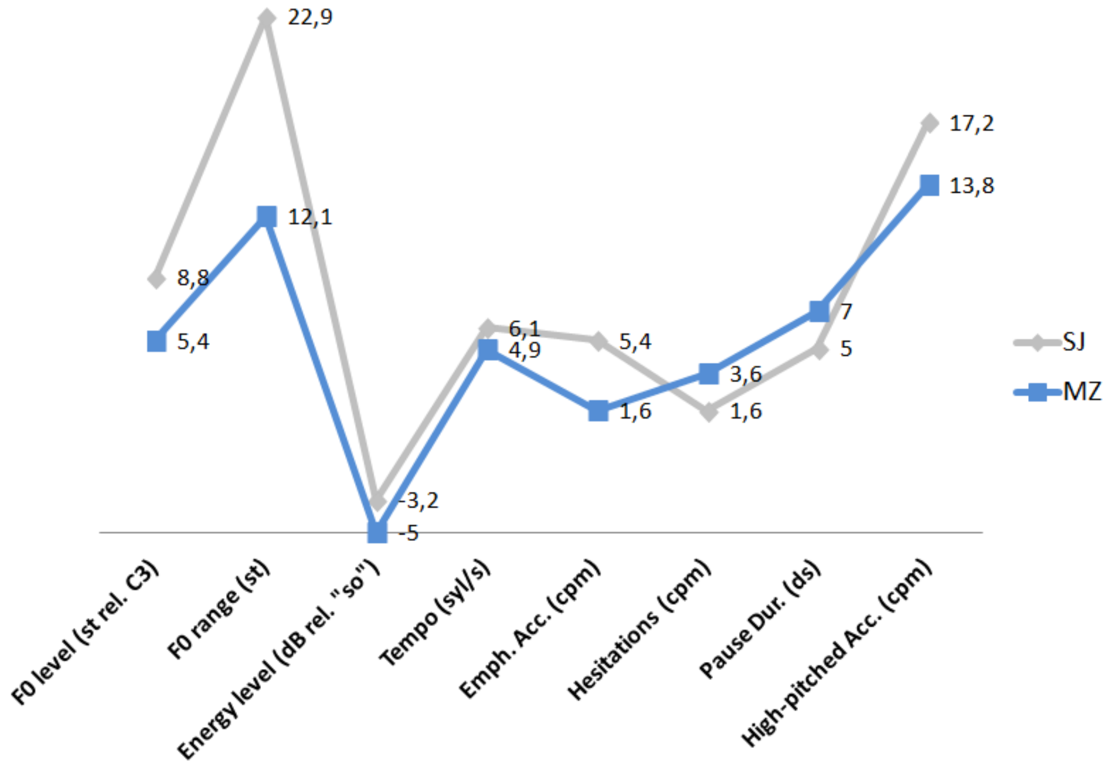

Figure 2: Prosodic features of the two CEOs (on average) in comparison regarding the parameters chosen

The mean parameter levels created in the robot speech stimuli to resemble Steve Jobs' and Mark Zuckerberg's speech characteristics are visualized in Figure 2; compared to the speech characteristics representing Mark Zuckerberg, we made the speech characteristics representing Steve Jobs faster and higher-pitched. Moreover, Jobs' speech characteristics had a higher acoustic energy level, a larger F0 range, shorter pauses between sentences or sentence sections, fewer hesitations (i.e. filled pauses), more emphatic accents and, in general, more high-pitched accents. With one exception, these created stimulus differences correspond to the parameter differences that characterize the real Jobs and Zuckerberg in public-speaking performances (Niebuhr et al. 2016a,b). The exception is tempo. While Mark Zuckerberg speaks generally fast, which leads to strong speech reduction and coarticulation phenomena (Niebuhr & Gonzalez to appear), which can hardly be modeled in speech synthesis, his speech is furthermore characterized by many disfluencies, which however are very hard to synthesize (Betz et al. 2015). We solved this problem by lengthening syllables with the same kind of frequency as his hesitation markers occur. Thus, the overall impression of his speech was expected to be similar. To confirm this hypothesis, following a recommendation from one of our reviewers, we checked the final pairs of stimuli with respect to their external validity. To that end, we took three randomly selected stimulus excerpts from the pool of Jobs-oriented and Zuckerberg-oriented stimuli. All excerpts were about 5 seconds long. Thirteen listeners were recruited (6 males, 7 females, mean age 23.1 years) who first watched the YouTube videos of the keynote speeches of Jobs and Zuckerberg, based on which the tone-of-voice profiles of the present stimuli were created. Then, the listeners were informed that they would hear six short speech syntheses excerpts produced by robots that tried to imitate either Jobs' or Zuckerberg's original tone-of-voice. Their task would be to decide spontaneously after each stimulus, whose tone-of-voice the robot had imitated. The results show that the Jobs-oriented stimuli were correctly identified as imitating Jobs' speech characteristics in 82.1% of the cases (i.e. 32 out of 39 decisions). For the Zuckerberg-oriented stimuli the correct identification rate was 69.2% (27 out of 39 decisions). A binominal test showed that the correct identification performances are in both conditions significantly above chance level (SJ: z=3.84, p<0.001, d=1.56; MZ: z=2.24, p=0.007, d=0.77) and do not differ statistically from each other, i.e. both sets of stimuli could be associated with their original speakers equally well.

During the creation of the robot speech stimuli, the manipulation procedure was conducted in an iterative way by which each parameter was successively adjusted. That is, we manipulated the respective parameter and saved the generated output in a wav file. Then, we analyzed the wav file acoustically and compared the mean value measured for the parameter to the corresponding reference mean value of

Steve Jobs or Mark Zuckerberg. Where the manipulated mean value deviated from its reference, we returned to the manipulation and made a further adjustment to the parameter. We continued with the increasingly fine-grained comparison-and-adjustment-loop until the targeted mean value was reached in the stimulus. For all parameters, the level of precision was set to one decimal place, partly because this is sufficient in terms of JNDs and partly for practical reasons (feasibility, limitations of digital-signal-processing tools). Recently, automatic, i.e. computer-based techniques, have been developed for a transfer of tone-of-voice characteristics from a source signal (like a specific phrase) to a target signal (another prosodic phrase), see, for example, Lorenzo-Trueba et al. (2015) and Skerry-Ryan et al. (2018). The use of these tools seems attractive for the purposes of the present study, however, since our manipulation not only involved the transfer of holistic acoustic settings from A to B, but also the creation and implementation of local linguistic units like filled pauses, hesitations and emphatic accents, we considered the manual iterative approximation of the target prosody in the stimuli via PSOLA the more suitable method. Another reason was that the lexical content i.e. the string of words of our target stimuli differed considerably from that of Jobs' and Zuckerberg's source signals.

The differences between the two robot voices are illustrated in Figure 3.[1] It shows *praat* displays of two similarly long sections from a Steve-Jobs and a Mark-Zuckerberg stimulus (used in Experiment III), with duration and pitch manipulations indicated. Besides the raised and extended pitch contour and the emphatic lengthening of accented syllables, it is obvious from Figure 3[2] that there are more syllables with fewer pauses and hesitations in the Jobs-oriented stimulus. Figure 3 also shows that we have taken into account and manipulated only a few of all those parameters that constitute a speaker's speech characteristics.

[1] The two video files can be found under the following links: https://youtu.be/4tuBEL9iQEk and https://youtu.be/RGXIX8WF7t4

[2] The two photographs in Figure 3 were created by *Justdoit709 [CC BY-SA 3.0], via Wikimedia Commons* (Steve Jobs) and by Dan Taylor for tech.eu. www.tech.eu (Mark Zuckerberg)

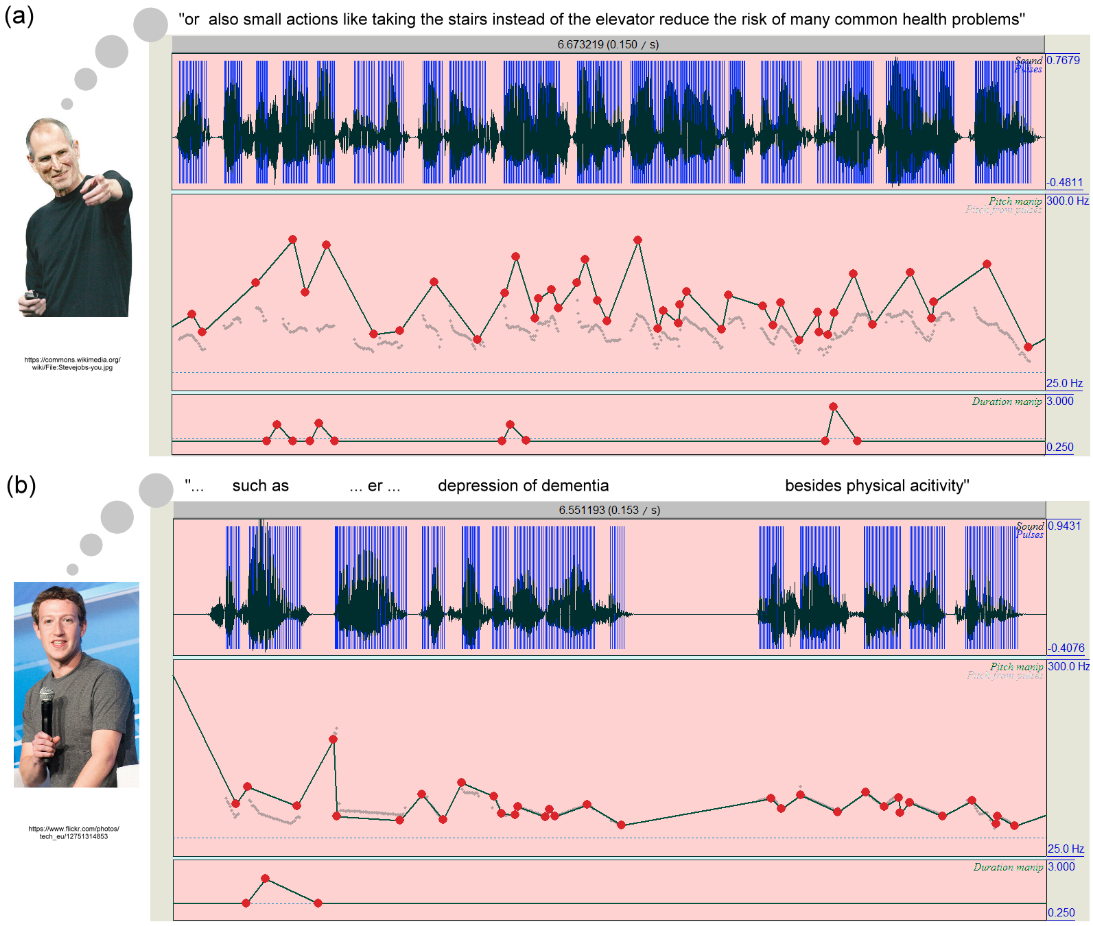

Figure 3: Praat displays of the two robot voices showing acoustic energy, number of syllables, disfluencies, accents and speech melodies

Based on the final, validity-checked pairs of stimuli, we conducted a series of human-robot interaction experiments that tested the effects of the stimuli's multi-parametric differences on human behavior and the robots' perceived personality traits. Note that despite the number of parameters that we manipulated in the present study, none of the stimuli we created really sounded like either Steve Jobs or Mark Zuckerberg. Hence, our results are definitely free of a speaker-identification bias. This is because a speaker's individual voice manifests itself acoustically primarily in the long-term energy distribution across the frequency spectrum that is determined by basic vocal-fold vibration characteristics (source signal) and the physiological filter characteristics of the vocal tract (acoustic resonator), see Dellwo et al. (2007).

Thus, we only changed the speech prosody, see Mannell 2007) with our manipulation, not the speaker's voice or voice quality itself. The voice used was always that of a male middle-aged Standard American-English speaker (neither Jobs or Zuckerberg) available from a free text-to-speech software. Accordingly, what we manipulated here concerns the two speakers’ different levels of expressiveness, but not to the identity of the speakers themselves.

Study I: Online Questionnaires

The first study tests the effects of speech melody in online questionnaires, i.e. in a situation in which the robot is not co-present, but participants only get to see a video of the robot. In the experiment, the robot greeted the participants and then asked them for a favor, in the voice with the characteristics of either SJ or MZ. The experiment was thus a between-subject online questionnaire. The persuasiveness of the robot’s speech melody was tested by asking the participants to fill out a longer rather than a shorter questionnaire (both questionnaires were in fact equally long). In particular, the robot says:

> *Hello, I'm the EZ-bot, thanks for taking the time to fill out this questionnaire – much appreciated! In fact there are two different questionnaires, one longer and one shorter, and I would be grateful if you could fill out the longer one, but thanks in any case!*

We took a video of the robot, the JD EZ-bot (see Fig.4), in which it moved slightly, and matched the video file with each of the audio files so that the robot's arm movements coincided with the speech accents. Participants had then the choice to click on the longer or on the shorter questionnaire.

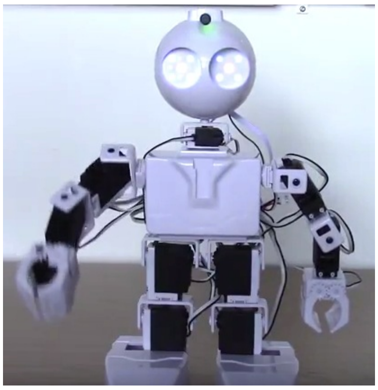

Figure 4: EZ-bot

Study I: Participants

209 completed the online questionnaire; due to the fact that the video did not run on tablets, we experienced a high drop-out rate on the second page. In addition, we removed 34 respondents from the analysis as their response time was shorter than the total duration of the video stimulus. We recruited participants through Prolific Academic and social media. Most respondents are native speakers of English, although there are smaller groups of native speakers of Danish, German and French. 117 participants are women, while 92 are men. The mean age of the participants is 33.3 (SD=12.8). 101 participants listened to the robot with Jobs speech characteristics,[3] and 108 participants heard the robot speaking to them with Zuckerberg's speech characteristics.[4]

Questionnaires

Following the video, respondents were asked to choose whether they want to complete a short or a longer questionnaire. Then they were asked to rate the robot. For the creation of the questionnaire, we relied on the measures used by Rosenberg & Hirschberg (2009) to study charismatic speech in humans, however, without the negative emotions "the speaker is intense, angry and accusatory" since they did not seem relevant to our scenarios. Instead, we added *engaging* to the list of attributes. In particular, we asked the participants to rate the robot on a scale from 1-7 on the following adjectives, where 1 corresponds to 'not at all' and 7 to 'very much':

- enthusiastic
- charming
- persuasive
- boring
- passionate
- convincing
- engaging

In addition, we asked to rate on a semantic differential scale from 1-5 whether, "overall, the interaction with the robot was...":

- fun ----- boring
- too long ----- too short
- exciting ----- tiring

Results

As is shown in Figure 5, overall 75.6% (i.e. 158) of our 209 valid participants took the long questionnaire. That is, less than one-fourth of our participants decided to ignore the robot's request and filled out the short questionnaire. However, the number of participants who took this easy way out was different between the two conditions. While in the condition with Jobs' speech characteristics only about every fifth participant (20.8% or 21) went for the short questionnaire, it was more than every fourth participant in Zuckerberg's speech characteristics condition (27.8% or 30) who chose to fill out the short questionnaire. However, this difference is not significant at $p<0.05$.

As for the subjective measures, we conducted a Multivariate Analysis of Variance (MANOVA) in order to compare, for each attribute, the rating scores received across all participants in the SJ and MZ conditions. Thus, Robot was the two-level fixed factor (Jobs vs Zuckerberg), and the 11 scales

[3] http://hri.sdu.dk/ez_SJ.html

[4] http://hri.sdu.dk/ez_MZ.html

represented the dependent variables, each with values between 1 and 5 or 7. Participant was included as a covariate.

The MANOVA showed a main effect of Robot (F[11,196]=2.24, p=0.014, η_p^2=0.12). The following scale-specific tests of between-subject effects revealed that this main effect relies mainly on four scale differences. Compared to the robot with Zuckerberg's speech characteristics, the robot using Jobs' speech characteristics sounded across participants more enthusiastic (F[1,206]=13.2, $p<0.001$, η_p^2=0.06), charming (F[1,206]=4.15, p=0.03, η_p^2=0.03), passionate (F[1,206]=9.62, p=0.002, η_p^2=0.05), and it was more fun to interact with (F[1,206]=4.52, p=0.02, η_p^2=0.03). Figure 5 provides a results summary of the rated scales. The covariate speaker was not significant.

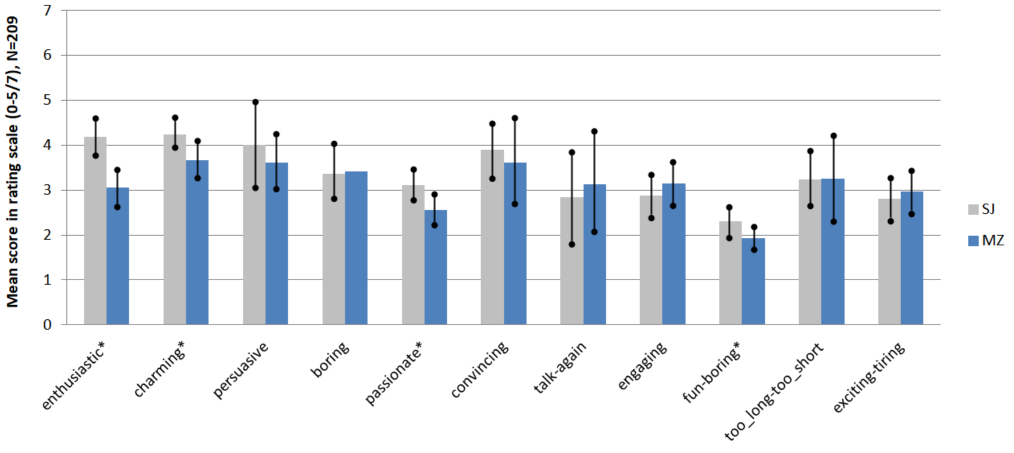

Figure 5: Results summary of rating-score differences of Experiment I, N=209. Asterisks indicate scales with significant differences between the SJ and MZ conditions.

Discussion

The analysis shows that while the robot's voice did not have a significant influence on participants' compliance with the robot's request, the robot's impressionistic evaluation of the robot's enthusiasm, charm and passion and of the extent with which they perceived the interaction as fun or boring was significantly affected by the robot's speaking style. All results point in the same direction, namely that the robot that uses the speech characteristics of Steve Jobs is rated more favorably, even though the difference is not significant for the behavioral variable.

Study II: Human-Robot Joint Tour Planning

For the second study, we replicated the scenario developed by Andrist et al. (2013, 2015), see Figure 6. In particular, we created speech stimuli for two Keepon robots, one with the prosodic characteristics of Steve Jobs, the other with the prosodic characteristics of Mark Zuckerberg. The order of presentation was changed so that the right robot talked like Steve Jobs in one condition and like Mark Zuckerberg in the other condition. Thus, our scenario differs from Andrist et al. (2013, 2015) only with respect to the order of presentation. While Andrist et al. switch the robot presenting, for instance, the expert versus non-expert stimuli during the interactions, this is obviously not possible when robots are speaking in different voices. We thus changed the order of presentation only between subjects, so that each participant heard each robot consistently speaking in one voice. In order to control for the order effect, we also investigated whether order of presentation influenced participants' choices (see the results section below). The stimuli (written text plus images) were tested in a pretest (n=32) whether they are largely equivalent; participants were asked to rate on a 7-point Likert scale to which place they would prefer to go, with 4 being neutral. The average value for the six stimuli is 4.5, with a standard deviation of 1.9.

Procedure

Participants' task was to plan a trip to Paris based on the suggestions made by the two Keepon robots. In particular, they saw two (very similar) pictures of churches, parks, catacombs, museums and fairgrounds while each robot delivered some background information about one of the choices. For example, the two robots advertised two bridges as follows:

This lovely old pull-bridge was built in 1745 when the first merchants moved into the area. It was initially made entirely of wood, which was replaced with cast iron in the 20th century. It connects the old business area to the medieval market area.

Built in the middle of the 18th century, the historical pull-bridge provides access into the renaissance business quarter and was once the most important connection to the upper part of town. Covered in bright white paint, it was also referred to as pont blanche – the white bridge.

The robots moved slightly when speaking; in particular, they turned towards the participant and made a little jump before they started talking and turned back to an initial state once they were done speaking. The speaking stimuli is counterbalanced with visual stimuli in order to negate any effects of differences between the images shown to participants. Overall, there were seven sightseeing stops to be made during the trip through Paris, which means that there were seven two-alternative forced-choice decisions for participants to be made in which they could either follow the Jobs-driven or the Zuckerberg-driven robot (or remain neutral).

Participants

Participants were students and staff at the University of Southern Denmark, campus Sonderborg. In total 15 participants took part in the study (8 men and 7 women) with an age range of 20-56.

Measures

The dependent variables are participants' choices of the sights recommended by the robots on the one hand and the questionnaire results on the other.

Behavioral Measures

The behavioral data used in this experiment was which robot's suggestions participants followed more. The order of presentation of the two robot voices was counter-balanced so that the voice with Mark Zuckerberg's speech characteristics appeared left for half of the participants and right for the other half.

Questionnaire

Participants rated the robots in comparison regarding the same characteristics as in Study I. Participants rated the robots on a scale with seven answer options, where the first answer option corresponds to the left Keepon and the seventh option corresponds to the right Keepon. Note that participants are not forced to decide for a particular Keepon but can also opt for no difference by ticking center value 4 on the rating scale.

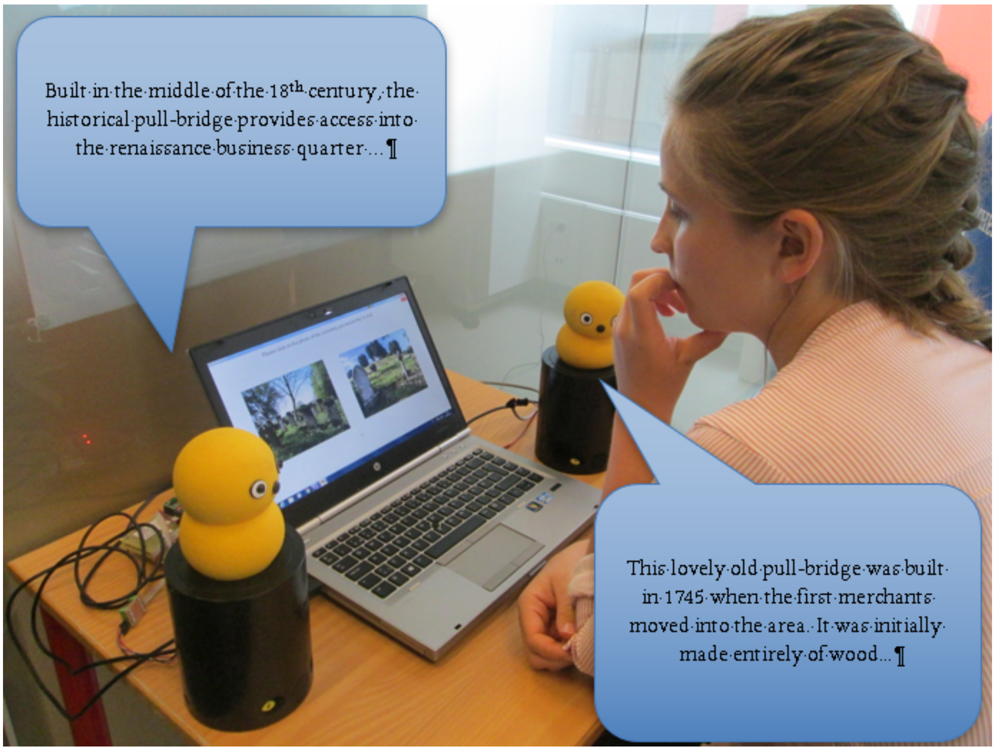
Built·in·the·middle·of·the·18th·century,·the·historical·pull-bridge·provides·access·into·the·renaissance·business·quarter·...¶
This·lovely·old·pull-bridge·was·built·in·1745·when·the·first·merchants·moved·into·the·area.·It·was·initially·made·entirely·of·wood...¶

Figure 6: Experimental set-up for Experiment II

The interaction was filmed by two cameras, and participants' choices were logged. After the experiment, participants filled out the questionnaire and were debriefed. They then received a bar of chocolate for their participation.

Results

The analysis of the persuasiveness of the two robots shows that the participants' choices of the sites to be included in their (imagined) trip to Paris was clearly influenced by the robots' speaking styles. As is shown in Figure 7, participants decided more often to visit the site that was advertised by the robot with SJ's speech characteristics. The overall differences across the two robot orders was 17.5 % in favor of the sites presented by the SJ-based robot (41 % vs. 59 % for the robot driven by Zuckerberg's vs Jobs' speech characteristics, respectively). A paired-samples t-test showed that this outcome is statistically significant with a large effect size in terms of Cohen's d (t[14]=2.11, p=0.043, d=1.1). Note that there was no significant effect for order (t[14]=0.25, p=0.802), i.e. whether the robot with Jobs' or Zuckerberg's speech characteristics was on the right or on the left side of the participant.

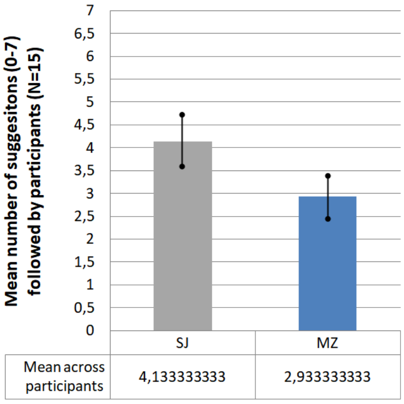

Figure 7: Results for Experiment II: behavioral choices for the two Keepons in terms of the mean number of suggestions (0-7) followed by the 15 participants when coming of the robot with Jobs' or Zuckerberg's speech characteristics.

Regarding the impressionistic evaluation of the two robots in the questionnaire after the experiment, we calculated a Multivariate Analysis of Variance (MANOVA) based on the 2-step fixed factor Robot (SJ's vs MZ's speech characteristics) and Participant as a covariate. The seven scales on which the two robots were rated represented the individual dependent variables, each of them with values between 1 and 7 that indicate whether the respective scale attribute was more applicable to the left or to the right robot. Of course, the rating data were cleaned with respect to participant-specific robot orders prior to the MANOVA. The MANOVA yielded a significant main effect of Robot (F[7,21]=3.10, p=0.021, η_p^2=0.51). Broken down by the individual 7 scales, we found that the robots were rated differently along four attributes. Compared to the robot using MZ's prosody, the robot that used the prosody of SJ was rated as more enthusiastic (F[1,27]=12.05, p=0.002, η_p^2=0.31), more charming (F[1,27]=8.10, p=0.008, η_p^2=0.23), more passionate (F[1,27]=8.62, p=0.007, η_p^2=0.24), and less boring (F[1,27]=5.14, p=0.032, η_p^2=0.16) by our participants.

The rating results are summarized in Figure 8, which shows that the questionnaire results all point in the same direction, such that the SJ-based robot received more positive ratings than the MZ-based robot for all traits besides *boring*.

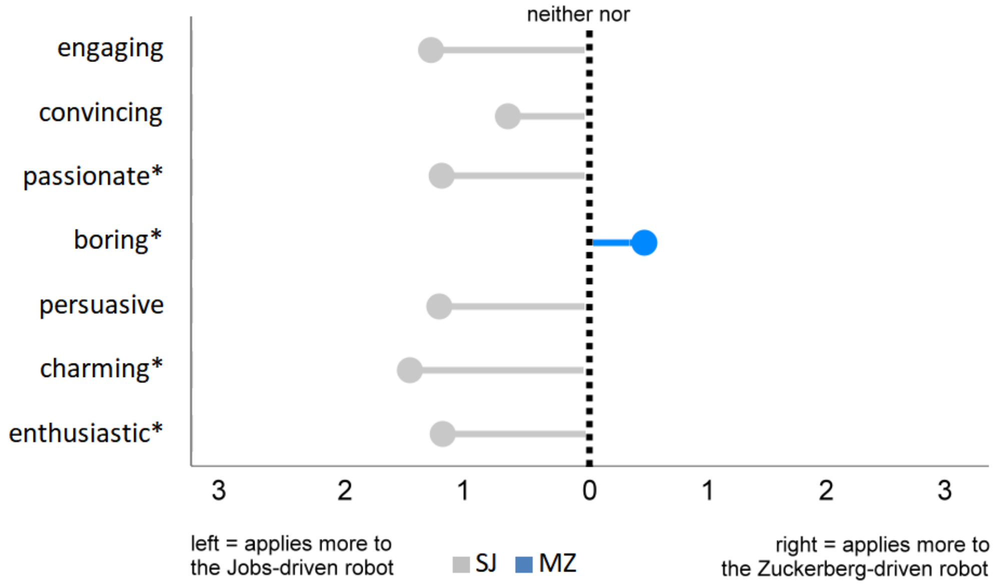

Figure 8: Results for Experiment II: Mean attribute ratings for the two Keepons across all 15 participants.

Discussion

The results show that the robot that uses SJ speech characteristics is consistently more persuasive than the robot that speaks with MZ's speech characteristics. Furthermore, also the impressionistic evaluation consistently suggests a more positive evaluation of the robot that speaks like Steve Jobs such that it is rated as more enthusiastic, charming, passionate and engaging. Just regarding the negative trait 'boring,' the MZ robot scores higher.

Study III: Service Robot

In the third study, the Care-O-bot3 (Graf et al. 2009; Jakobs & Graf 2012, see Fig. 1) was used to lecture participants about a healthier life-style (in particular, regular health checks, sugar intake and taking the stairs) and then to offer participants to measure their blood pressure. The text the robot produced was:

> *Hello, I hope you are doing fine!*
> *I wanted to tell you a bit about latest findings concerning the effects of physical activity and a healthy life style. It is relatively well known that physical activity such as doing sports regularly or also small actions like taking the stairs instead of the elevator reduce the risk of many common health problems, like cardio-vascular disease, diabetes, osteoporosis and even cancer.*

What is lesser known is that physical activity and a healthier life style also contribute to psychosocial health. For instance, regular physical activity has a positive influence on cognitive and affective disorders, such as depression or dementia.
Besides physical activity, a healthier life style includes reducing the amount of sugar intake and regular check-ups on how you are doing health-wise. If you wish, I can measure your blood pressure. Would you like me to do that?
Could you put your arm under my sensor?
Thank you, that's it already. I wish you a very good day – and stay healthy!

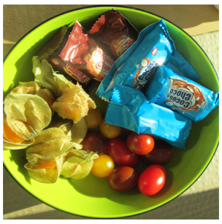

Figure 9: Bowl with fruit and sweets

Procedure

Participants were greeted in the hallway and asked to fill out a questionnaire with demographic information. They were then introduced to the Care-O-bot "and this is the Care-O-bot!", after which the robot started to speak. After the question whether the participant wants his or her blood pressure measured, the robot paused. If participants agreed, the robot moved its arm from its back to the front and instructed the participants to put their arm under its optical sensor. After a few seconds, the robot thanked the participant (see the robot's text above).

Participants were then asked to fill out a questionnaire; the questionnaire used is again the same as in the other two studies. In addition, participants were also asked, again on a 5-point scale ranging from 'very much' to 'not at all', to rate the degree to which the interaction with the robot was exciting or tiring, fun or boring, and too long or too short. Finally, participants were asked the extent to which they wanted to talk to the robot again. Towards the end of the questionnaire, they were offered a 'refreshment' from a bowl which was half filled with sweets and half filled with fruit (cherry tomatoes and physalis, see Fig. 9). Participants were then thanked and debriefed (especially concerning the fact that there was no real medical check-up) and invited to take part in a second experiment (not reported on here). Those who accepted this invitation were sent up two floors where the other experiment took place. For these participants, we noted down who took the stairs and who took the elevator.

Participants

31 students (9 women, 22 men) from the University of Southern Denmark, campus Odense, participated in the experiment. Their ages range from 19 to 33. Most participants (64.5 %) are native speakers of Danish while others are native speakers of Portuguese, Arabic, Dutch, Spanish, and Faroese.

Results

Participants in Experiment III only heard the robot speaking with either the prosody of SJ or of MZ. Therefore, we used independent-samples tests in order to statistically analyze our results. As some of the results concerning the 5-point rating scales turned out in F-tests to be not normally distributed, we used non-parametric statistics to analyze the data statistically. Effect sizes (η^2) are reported together with test statistics below. As is shown in Figure 10, three of the eleven scales that were included in the questionnaire yielded a significant difference. Compared to the robot using Zuckerberg's speech characteristics, the robot using Jobs' speech characteristics was rated by participants to sound more enthusiastic (U[15,16]=32.5, $p<0.001$, $\eta^2=0.39$), more charming (U[15,16]=62.5, $p=0.02$, $\eta^2=0.17$), and more passionate (U[15,16]=71.0, $p=0.05$, $\eta^2=0.12$).

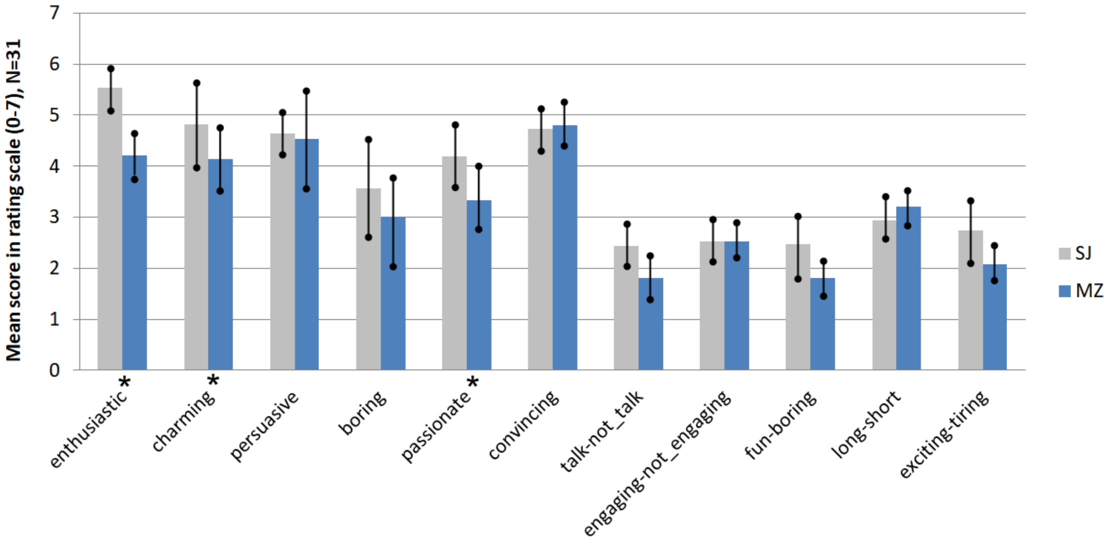

Figure 10: Results summary of rating scales used in Experiment III, N=31. Asterisks indicate scales with significant differences between the SJ and MZ conditions.

As for the behavioral data, we find that in the case of MZ's voice, only one third, i.e. 33%, of the participants went for a fruit, whereas the vast majority of 67% took a sweet from bowl of refreshments offered to them. In contrast, for SJ's voice 11% more, i.e. 44% of the participants chose a fruit (the difference is, however, not significant at p<0.05).

Furthermore, all participants in both conditions wanted the robot to measure their blood pressure, and they all took the stairs after the experiment. As for the other behavioral variables, we were thus not able to find any behavioral differences.

Discussion

Like in the previous studies, the robot that used Job's speech characteristics was rated more positively with respect to traits that are related to charisma and expressive speech, i.e. enthusiasm, charm, and passion. Differences on other scales fit in with this pattern, for example, that participants preferred talking to the Jobs-driven robot and that it was more fun and exciting to talk to it. However, these differences did not become significant. The same applies to the behavioral differences, i.e. the 11% increase in fruit consumption over chocolate consumption if the advice for a healthier lifestyle came from the robot with Jobs' speech characteristics.

General Discussion

Our results show that prosodic characteristics of robot speech such as the speech melody may influence both the persuasiveness of the respective robot and its impressionistic evaluation. Regarding robot persuasiveness, given that behavioral choices always depend on the circumstances, such as, according to participants' own explanations, time of day, amount of stress, curiosity or the fact that they had brought an apple with them already, just to name a few, the effects of people favoring the voice characteristics of Steve Jobs, i.e. 11% in Experiment III to more than 17% in Experiment II, are actually quite considerable. That a healthier life is more often reflected in choosing the healthy food option (fruit) when it is advertised by SJ's prosody is consistent with what was expected. Furthermore, the tendency for the healthier snack is perfectly in line with the behavioral changes in the other two experiments. Together, the three experiments indicate a behavioral effect across activity types: The behavioral measure in Experiment I was a request, in Experiment II it was a suggestion, and in Experiment III, it was an indirect recommendation. In all three cases, the success rate was higher (i.e. the robot was more persuasive) in combination with SJ's than in combination with MZ's speech characteristics. Only the behavioral effect of Experiment II became statistically significant, though. It is not unlikely that this results difference between Experiment II and the other two experiments is due to the fact that the behavioral measure in Experiment II consists of the average over seven decisions (i.e. the seven sights people decided to see). In the other two experiments, we are dealing with single binary decisions (i.e. whether or not to fill out the long questionnaire and whether to choose fruit over sweets) that do not allow a differentiated, quantitative assessment of individual participants' behavior. It is obvious that this reduces the sensitivity of the dependent variable. Thus, future studies should aim to design experiments in such a way that they measure quantifiable rather than binary behavioral differences between speech conditions. Note that the behavioral differences triggered by the Jobs-oriented and Zuckerberg-oriented stimuli in Experiments I and III are also significant if they are pooled across the two experiments ($\chi^2[1]=3.5$, $p=0.03$, $d=0.49$), one-tailed test with reference to the outcome of Experiment II). That is, in combination, the behavioral data of Experiments I and III are consistent with those of Experiment II in that Jobs' speech characteristics had a significantly higher persuasive power than Zuckerberg's speech characteristics in that it made significantly more participants follow the robot's request (long instead of short questionnaire) and implicit advice (fruit over chocolate).

Similarly, also the impressions the robots in all three studies made on participants clearly point in the same direction. These findings suggest that speech characteristics should be taken into account in robot speech synthesis in the future.

Especially with respect to the small sample sizes of Experiments II and III, we did not test for gender-specific effects in the present paper, knowing, however, that listener gender generally plays a role in detecting and assessing the emotions of speakers (Lambrecht et al. 2014), including those related to speaker charisma. Informal observations in our data suggest that there are gender-specific differences of two different kinds in our results. First, in terms of the robots' traits, male ratings seem to differ more strongly than female ratings across the three experiments. Second, while the male listeners seem to produce the strongest rating difference between the Jobs and Zuckerberg conditions along the robots' emotional qualities (e.g., charm and passion), the strongest rating differences among female listeners seem to concern interactional qualities (e.g., persuasion, fun in interaction, too long/short interaction time). Further studies will pursue these potential gender-differences in separate experiments, building on the pilot study of Novák-Tót et al. (2017). In any event, even if female and male participants may have differed regarding the interpretation of particular traits, they still consistently judged the robots with Steve Jobs' speech characteristics more favorably.

Another future study will address whether our manipulations resulted in differences in the quality of the PSOLA resyntheses between the Jobs and Zuckerberg conditions, which we believe to be highly unlikely since the manipulation of the Jobs-like speech files required more adjustments than the resynthesized Zuckerberg-oriented stimuli. Nevertheless, subsequent experiments will include rating scales that specifically address the issues of naturalness, intelligibility and sound quality.

Conclusion and Future Work

To conclude, robots that use speech characteristics from Steve Jobs have been found to be more persuasive and to be rated more favorably than robots that use speech characteristics from another powerful CEO, Mark Zuckerberg. Thus, simply to copy some general human speech characteristics will not be sufficient when the aim is to make robots influential and well received. While we have not tested the robots' influence when using unmanipulated speech based on public domain text-to-speech systems, previous work indicates that this will be judged even worse (e.g. Walters et al. 2008; Nass & Brave 2005). These findings indicate that an attention to robot speech characteristics is crucial.

While our results suggest that human-robot interactions should be designed with robot speech in mind, the necessary technologies to do so are not there yet to the extent required to synthesize persuasive, context-aware, charismatic robot speech automatically, nor are the linguistic definitions of what characteristics are useful in what situation or methods for the evaluation of such definitions.

Compliance with Ethical Standards:

The data gathered for the studies reported on were elicited according to Danish legislation. Participants were asked before the experiments whether they agree to being videotaped and whether they agree to their pictures being taken and used in publications and presentations. We are using data only from participants who have consented in writing that they agree to this. Participants were informed that they can stop the experiments at any time. After the experiments, participants were debriefed, and it was explained to them in writing that no actual medical data had been obtained. All data were stored and handled anonymously.

Funding:

This study was partially funded by the Danish Innovation Fund (projects patient@home and *Smooth*).

Acknowledgements:

We wish to thank Janni Jensen and Nathalie Schümchen for their help during the experiments.

Conflict of Interest:

The authors declare that they have no conflict of interest.

References

[1] Andrist, S., Spannan, E., and Mutlu, B. (2013). Rhetorical Robots: Making Robots More Effective Speakers Using Linguistic Cues of Expertise. In Proceedings of the 8th ACM/IEEE International Conference on Human Robot Interaction (HRI '13). IEEE Press. Piscataway, NJ, USA. 341-348.

[2] Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., and Sakr, M. (2015). Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15). ACM. New York, NY, USA. 157-164.

[3] Bainbridge, W.A., Hart, J.W., Kim, E.S. & Scasselati, B. 2010. The Benefit of Interactions with Physically Present Robots over Video-Displayed Agents. International Journal of Social Robotics 1-2.

[4] Berger, S., Niebuhr, O., and Peters, B. (2017). Winning Over an Audience – A Perception-based Analysis of Prosodic Features of Charismatic Speech. Proc. 43rd Annual Meeting of the German Acoustical Society (DAGA), Kiel Germany, 1-4.

[5] Betz, Simon, Wagner, Petra & Schlangen, David (2015). Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. *Interspeech* 2015.

[6] Biadsy, F., Hirschberg, J., Rosenberg, A. and Dakka, W. (2007). Comparing American and Palestinian perceptions of charisma using acoustic-prosodic and lexical analysis. Proceedings of 8th Interspeech Conference, Antwerp, Belgium, 2221-2224.

[7] Boersma, P. (2001). Praat: A system for doing phonetics by computer. Glot International 4, 341-345.

[8] Campbell, N. and Mokhtari, P. (2003). Voice quality - The 4th prosodic dimension. Proc. 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2417-2420.

[9] Crumpton, Joe & Bethel, Cindy 2016. A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. International Journal of Social Robotics (2016) 8:271–285.

[10] Chidambaram, Vijay, Chiang, Yueh-Hsuan and Bilge Mutlu (2012). Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues. HRI'12, March 5–8, 2012, Boston, MA.

[11] Dellwo, V., Huckvale, M., and Ashby, M. (2007). How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In Müller, C. (Ed.), Speaker Classification I, pp. 1-20. New York: Springer.

[12] Dutoit, T. (2013) An Introduction to Text-To-Speech Synthesis. Dordrecht: Kluwer.

[13] Gélinas-Chebat, C., Chebat, J.C., and Vaninski, A. (1996). Voice and advertising: effects of intonation and intensityof voice on source credibility, attitudes toward the advertising service and the intent to buy. Perceptual and Motor skills 83: 243-262.

[14] Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2003) (pp. 55–60). Los Alamitos, CA: IEEE.

[15] Graf, B., Reiser, U., Hagele, M., Mauz, J. and P. Klein (2009) Robotic home assistant care-o-bot 3-product vision and innovation platform. IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO) 2009, pp. 139–144.

[16] Gregory, R. L. (1997). Eye and brain – The psychology of seeing. Oxford: Oxford University Press.

[17] Grigore, Elena Corina, Pereira, Andre, Zhou, Ian, Wang, David, and Scassellati, Brian (2016). Talk to Me: Verbal Communication Improves Perceptions of Friendship and Social Presence in Human-Robot Interaction. Proceedings of the 2016 Intelligent Virtual Agents Conference, IVA'16.

[18] Heracleos, Loizos and Klaering, Laura Alexa (2014). Charismatic Leadership and Rhetorical Competence. An Analysis of Steve Jobs's Rhetoric. Group and Organization Management 39, 131–161.

[19] In, Jiyoung and Han, Jeonghye (2015). The Prosodic Conditions in Robot's TTS for Children as Beginners in English Learning. Indian Journal of Science and Technology 8; 55: 48-51.

[20] Jacobs, Theo & Graf, Birgit (2012). Practical Evaluation of Service Robots for Support and Routine Tasks in an Elderly Care Facility. In: IEEE Workshop on Advanced Robotics and its Social Impacts - ARSO 2012: 21-23 May 2012, Munich, Germany. Piscataway, NJ: IEEE, 2012, S. 46-49

[21] Jung, Malte (2017). Affective Grounding in Human-Robot Interaction. Proceedings of HRI'17, Vienna

[22] Kanda, T., Miyashita, T., Osada, T., Haikawa, Y. and Ishiguro, H. 2008. Analysis of Humanoid Appearances inHuman-Robot Interaction. IEEE Transactions on Robotics 24,(3). 725-735.

[23] Ladd, D.R. 2014. Simultaneous Structure in Phonology. Oxford. OUP.

[24] Lambrecht, L., Kreifelts, B., Wildgruber, D. (2014) Gender differences in emotion recognition: Impact of sensory modality and emotional category, Cognition and Emotion, 28:3, 452-469, DOI: 10.1080/02699931.2013.837378

[25] Landgraf, R. (2014). Are you serious? Irony and the perception of emphatic intensification. Proc. 4th International Symposium on Tonal Aspects of Languages (TAL),Nijmegen, The Netherlands, 91-94.

[26] Leyzberg D, Avrunin E, Liu J, Scassellati B (2011) Robots that express emotion elicit better human teaching. In: Proceedings of the 6thACM/IEEE international conference on human–robot interaction (HRI), pp 347–354.

[27] Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., & Montero, J. M. (2015). Emotion transplantation through adaptation in HMM-based speech synthesis. Computer Speech & Language, 34(1), 292-307.

[28] Mannell, R. (2017). Phonetics and Phonology. Introduction to prosody: Theories and models. URL: http://clas.mq.edu.au/speech/phonetics/phonology/intonation/index.html

[29] Moore, B.J. (2013) An Introduction to the Psychology of Hearing. Leiden: Brill

[30] Mutlu, Bilge (2011). Designing Embodied Cues for Dialog with Robots. AI Magazine 32, 4: 17-30.

[31] Nass, Clifford & Brave, Scott (2015). Wired for Speech. How Voice Activates and Advances the Human-ComputerRelationship. The MIT Press, Cambridge, MA.

[32] Niebuhr, O. & S. Gonzalez (to appear). Do sound segments contribute to sounding charismatic? Evidence from acoustic vowel space analyses of Steve Jobs and Mark Zuckerberg. International Journal of Acoustics and Vibration 24/2

[33] Niebuhr, O., R. Skarnitzl & L. Tylečková (2018). The acoustic fingerprint of a charismatic voice - Initial evidence from correlations between long-term spectral features and listener ratings. Proc. 9th International Conference of Speech Prosody, Poznan, Poland, 359-364.

[34] Niebuhr, O., S. Tegtmeier, & A. Brem (2017). Advancing research and practice in entrepreneurship through speech analysis – From descriptive rhetorical terms to phonetically informed acoustic charisma metrics. Journal of Speech Science 6, 3-26.

[35] Niebuhr, O., H. Reetz, J. Barnes & A. Yu (2019). Fundamental aspects in the perception of f0. In: C. Gussenhoven, A. Chen (Eds), The handbook of prosody. Oxford: Oxford University Press.

[36] Niebuhr. O. (2017). Clear Speech - Mere Speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. Proc. 18th Interspeech Conference, Stockholm, Sweden, 1-5.

[37] Niebuhr, O., Voße, J., and Brem, Am (2016a). What makes a charismatic speaker? A computer-based acousticprosodic analysis of Steve Jobs tone of voice. Computers in Human Behavior 64: 366-382.

[38] Niebuhr, O., Brem, A., and Nowak-Tót, E. (2016b). Prosodic constructions of charisma in business speeches – A contrastive acoustic analysis of Steve Jobs and Mark Zuckerberg. Proc. 8th International Conference of Speech Prosody, Boston, USA, 1-2.

[39] Nienhuis, M. (2009). Prosodic Correlates of Rhetorical Appeal: Voice Wave Analysis of Ethos, Pathos and Logos. Proc. 10th Twente Student Conference, Twente, The Netherlands, 1-7.

[40] Nishio, S, Ogawa, K., Kanakogi, Y., Itakura, S. and Ishiguro, H. 2012. Do Robot Appearance and Speech Affect People's Attitude? Evaluation through the Ultimatum Game. Proceedings of Ro-Man 2012, Paris.

[41] Novák-Tót, E., Niebuhr, O., and Chen, A.(2017). A gender bias in the acoustic-melodic features of charismatic speech? Proc. 18th Interspeech Conference, Stockholm, Sweden, 1-5.

[42] Pejčić, A. (2014). Intonational Characteristics of Persuasiveness in English and Serbian. Noveaux Cahiers de Linguistique Francaise 31, 141-151.

[43] Powers, A. and Kiesler, S. 2006. The Advisor Robot: Tracing People's Mental Model from a Robot's Physical attributes. Proceedings of the Human-Robot Interaction Conference HRI'06, Salt Lake City, Utah, USA.

[44] Read, Robin & Belpaeme, Tony (2010) Interpreting non-linguistic utterances by robots: Studying the influence of physical appearance. In: Proceedings of the 3rd international workshop on affective interaction in natural environments (AFFINE). ACM, New York, pp 65–70.

[45] Read, Robin & Belpaeme, Tony (2014) Situational context directs how people affectively interpret robotic non-linguistic utterances. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 41–48.

[46] Rosenberg, A. and Hirschberg, J. (2005). Acoustic/Prosodic and Lexical Correlates of Charismatic Speech. Proc. 6th Interspeech Conference, Lisboa, Portugal, 513-516.

[47] Rosenberg, A. and Hirschberg, J. (2009). Charisma perception from text and speech. Speech Communication 51:640-655.

[48] Selting, Margret (1994). Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation. Journal of Pragmatics 22, 375-408

[49] Rudzicz, F., Wang, R., Begum, M. and Mihailidis, A. (2015). Speech interaction with personal assistive robots supporting aging-at-home for individuals with Alzheimer's disease. ACM Transactions on Accessible Computing, 7(2).1–22.

[50] Signorello, R. and Demolin, D. (2013). The physiological use of the charismatic voice in Political speech. Proc. 14th Interspeech Conference, Lyon, France, 987-991.

[51] Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. arXiv preprint arXiv:1803.09047.

[52] Stern, S.E., Mullennix, J.W., Yaroslavsky, I. (2006). Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. International Journal of Human-Computer Interaction 64, 43-52.

[53] Strait, M., Canning, C. and Scheutz, M. 2014. Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. Proceedings of HRI'14, Bielefeld, Germany.

[54] Schuller, B. et al. (2015). A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge. Computer Speech and Language 29: 100-131.

[55] Stibbard, R. M. (2001). Vocal Expression of Emotions in Non-laboratory Speech: An Investigation of the Reading/Leeds Emotion in Speech Project Annotation Data. Unpublished PhD thesis. University of Reading, UK.

[56] Taylor, P. (2009). Text-to-speech synthesis. Cambridge: Cambridge University Press.

[57] Theune, M., Klabbers, E., Odijk, J., de Pijper, J.R. and Krahmer, E. (2001). From Data to Speech: A general approach. Natural Language Engineering 7 (1): 47-86

[58] Tielman, M, Neerincx M, Meyer JJ, Looije R (2014) Adaptive emotional expression in robot-child interaction. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction (HRI). ACM, New York, pp 407–414.

[59] Walters, M.L., Dyrdal, D.D., Koay, K.L., Dautenhahn, K. and te Boeckhorst, R. (2008). Human Approach Distances to a Mechanical-Looking Robot with Different Voice Styles. Proceedings of Ro-Man 2008, Munich, Germany.