Behaviour & Information Technology, 2013
Vol. 32, No. 2, 117–131, http://dx.doi.org/10.1080/0144929X.2012.702285

Taylor & Francis
Taylor & Francis Group

Subjective responses to synthesised speech with lexical emotional content: the effect of the naturalness of the synthetic voice

Mirja Ilves* and Veikko Surakka

Research Group for Emotions, Sociality, and Computing, Tampere Unit for Computer-Human Interaction (TAUCHI), School of Information Sciences, University of Tampere, Kanslerinrinne 1, FI-33014 University of Tampere, Finland

(*Received 5 September 2011; final version received 7 June 2012*)

This study aimed to investigate how the degree of naturalness and lexical emotional content of synthesised speech affects the subjective ratings of emotional experiences and how the naturalness of the voice affects the ratings of voice quality. Twenty-four participants listened to a set of affective words produced by three different speech synthesis techniques: formant synthesis, diphone synthesis and unit selection synthesis. The participants' task was to rate their experiences evoked by the speech samples using three emotion-related bipolar scales for valence, arousal and approachability. The pleasantness, naturalness and clarity of the voices were also rated. The results showed that the affective words produced by the synthesisers evoked congruent emotion-related ratings in the participants. The ratings of the experienced valence and approachability were statistically significantly stronger when the affective words were produced by the more humanlike voices as compared to the more machinelike voice. The more humanlike voices were also rated as statistically significantly more natural, pleasant and clear than the less humanlike voice. Thus, our findings suggest that even machinelike voices can be used to communicate affective messages but that increasing the level of naturalness enhances positive feelings about synthetic voices and strengthens emotional communication between computers and humans.

Keywords: emotion; synthesised speech; naturalness

1. Introduction

A significant number of studies have shown that emotions affect human behaviour, including cognitive processes, in various ways (e.g. Laird *et al*. 1982, Isen *et al*. 1987, Estrada *et al*. 1994, Fredrickson and Branigan 2005, Spering *et al*. 2005, Minnema and Knowlton 2008, Werner *et al*. 2009). Further, it has been shown that emotional behaviour during human–human communication has an influence on the quality of interaction (Zajonc 1980) and on the emotions that are evoked in the participants of interaction (Surakka and Hietanen 1998).

The importance of emotions for human behaviour has been recognised as an essential part of human–computer interaction (HCI) research as well. Affective computing, the research area emphasising the importance of emotions in HCI, essentially relates to all processes (i.e. experience, behaviour and physiology) that can arise while interacting with computers. One of the specific goals in affective computing is to build emotionally intelligent systems that can recognise a user's emotions (e.g. from the user's face, voice or physiology) during interaction and then respond in a way that takes emotions into account (Picard 1995, Klein *et al*. 2002, Conati and Maclaren 2009). Emotion detection is very challenging, and in spite of the progress made in this area, it is still in its infancy (Calvo and D'Mello 2010). Another way to increase the emotionality of interaction is to develop the emotional expressiveness of computers in order to affect the emotional state of a computer user. For example, positively changing the emotional mind-set of the user can result in the user's improved performance (Aula and Surakka 2002, Partala and Surakka 2004). One significant component of emotions in HCI relates to emotional experiences evoked during the interaction. Different approaches have been employed to analyse emotional experiences in HCI. One is related to a wider concept called the 'user experience', which refers to the overall experience arising from the interaction between a human and a piece of technology. It is widely accepted that emotions are a central part of this experience (Hassenzahl and Tractinsky 2006, Law and van Schaik 2010). The study of human emotions has a long tradition, and it has been reasonable to measure users' emotional experiences using the methods and theories developed in core emotion research.

*Corresponding author. Email: mirja.ilves@sis.uta.fi

In order to add emotional expressiveness for HCI, computers can be programmed to express emotions through facial expressions, the gestures of animated agents and synthesised voices. There is evidence that users can accurately recognise emotions from the facial expressions of an embodied agent/avatar (e.g. Bartneck 2001, Ku *et al*. 2005) or from synthesised voice expressions (e.g. Cahn 1990, Murray and Arnot 1995, Iida *et al*. 2003). However, no common understanding exists how these synthetic emotional expressions influence humans. Previous studies have mainly explored the influence of synthetic emotions on users' attitudes, perceptions, behaviours and performance (see reviews by Dehn and van Mulken 2000, Beale and Creed 2009), and the findings have been somewhat ambiguous. While some studies have not found any significant effects of emotional agents, others have found that emotional agents can positively influence users. For example, an animated computer agent expressing empathy may decrease a user's stress and frustration (Prendinger *et al*. 2005) and evoke positive opinions regarding the agent's likeability and trustworthiness (Brave *et al*. 2005). Beale and Creed (2009) highlighted a number of questions that should be studied in order to develop a deeper understanding of the effects of emotions in human–agent interaction. One such question addresses whether it is possible to evoke the same emotion in a user as an agent is expressing (i.e. can the emotions expressed by a computer agent be contagious?). Certain findings suggest that this may be the case. For example, an avatar's happy expressions have been shown to evoke congruent facial muscle reactions in viewers (Weyers *et al*. 2006). It has also been found that synthetic speech with emotionally positive content can evoke congruent emotional experiences and result in improved human cognitive performance during computerised problem-solving tasks (Aula and Surakka 2002, Partala and Surakka 2004). More research is still needed to deepen and broaden the understanding of the emotional effects of synthesised emotions on humans, and this article aims to study how humans' emotional experiences are affected by the synthetically produced verbal emotional expressions of a computer system.

Several studies conducted during the past two decades have shown that voice technology provides a promising tool for realistic social interaction (Nass and Brave 2005). There is evidence that different voices presented by a computer are treated as if they are distinct social actors, even when the voices are presented by the same computer (Nass *et al*. 1994). Further, people recognise personality cues from synthesised voice and are attracted to voices that match their own personalities (Nass and Lee 2001). Overall, the research has shown that technology-generated speech evokes the same kind of responses and behaviour as speech coming from an actual person (Nass and Brave 2005). These findings demonstrate the Computers are Social Actors (CASA) paradigm, which means that people apply social rules towards interactive technologies (Reeves and Nass 1996). The use of speech in HCI can also have other types of advantages. Current interfaces are usually overloaded with visual information, so the use of speech has some advantages over textual or other visual information when communicating messages from a computer to a user. According to multiple resource theory (Wickens 2002), not only the difficulty of tasks but also the qualitative demands of tasks affect dual task performance. That is, dividing attention between the visual and auditory channels is usually easier than dividing it between two visual or two auditory channels. In addition, Alais *et al*. (2006) found evidence that audition and vision are under separate attentional control, at least to some extent. Thus, when a user is carrying out some visual task, speech messages likely disturb visual processing less than visual messages. For these reasons, computer-generated speech offers great potential for providing, for example, feedback or encouragement to computer users. Overall, speech research in the field of HCI has concentrated, for example, on studying the above-mentioned social responses, speech-recognition systems (e.g. Dai *et al*. 2005, Shi and Zhou 2009) and emotional expressivity in synthetic speech (see Schröder 2009), but studies on how speech messages affect the human emotional system are lacking.

In speech, emotions can be communicated through the prosodic features of the voice. There is evidence that discrete emotional states – such as happiness, anger, fear and sadness – cause divergent changes in speech prosody, especially in pitch level and range, speech rate and intensity (Kappas *et al*. 1991, Murray and Arnot 1993). Besides (or instead of) prosodic cues, emotions can also be communicated through the lexical content of speech. While some emotions have been modelled successfully by synthetic speech, there is still a long way to go before emotional speech synthesis is applicable in real-life settings. No exhaustive knowledge of the acoustic parameters relating to emotions exists, and controlling these parameters in a high-quality synthesis is a technological challenge (Schröder 2009). In addition, researchers are concentrated on the modelling of extreme emotional states, although the analysis of spontaneous speech has shown that the intensity of emotions expressed in speech is often weak or moderate and that expressions tend to be more mixed than discrete (Cowie and Cornelius 2003). Thus, the systems that express less intense emotional expressions would often be more suitable for real-life applications (Schröder 2009).

While emotional speech synthesis is developing, it is also worth studying how human emotions are affected by the lexical emotional content of speech. Our approach of studying speech in HCI is to investigate how the verbal content of messages affects the emotional experiences of humans. There is already some evidence that hearing synthetically produced sentences that evaluate a listener's emotional state, describe a computer's emotional state or provide emotional statements of a listener evokes emotional reactions in listeners (Ilves and Surakka 2009, Ilves *et al.* 2011). This happened even when there was no interaction between a human and a computer and the statements were purely random. These results suggest that it may be possible to regulate a user's mental state by the emotional statements. Thus, for example, positive language could be utilised in order to evoke positive feelings when interacting with different speech interfaces such as automated spoken dialogue systems.

Synthesised voices can be created by using different techniques. Two main techniques have been rule based or formant synthesis and concatenative synthesis (Dutoit 1997). Formant synthesis is created by using a mathematical model of speech sounds, so it does not apply human speech recordings at runtime. One weakness of formant synthesis is that its output sounds quite unnatural. The strengths of formant synthesis are in its relatively low computational requirements and high degree of control over acoustic parameters. Concatenative synthesis is reconstructed of the recordings of real human speech. Concatenated units can be, for example, phones, diphones, words or even sentences. The longer the units, the more natural the voice sounds and the more computation resources required. The two main subtypes of concatenative synthesis are diphone synthesis and unit selection synthesis. The database of diphone synthesis consists of all the diphones (transition between two phones) occurring in a language and the size of the database is relatively small. In contrast, unit selection synthesis is based on a large database in which the most appropriate units are selected and concatenated together. The output of this technique is usually perceived as the most natural by listeners (Schröder 2009). Recently, also hidden Markov models (HMM) based speech synthesis systems have emerged. In HMM-based speech synthesis method, HMMs are trained from the databases of natural speech and it generates speech waveforms from HMMs themselves. The output is not as high quality as the output of unit selection, but the voice characteristics of HMM-based synthesis systems can be modified by transforming HMM parameters. Thus, speaking style can be adjusted without the recordings of very large databases.

Thus, synthesised voices can be more or less anthropomorphic – they may vary from machinelike to human sounding. Originally, anthropomorphism referred to the tendency to attribute human characteristics to nonhuman things or events (Guthrie 1993). In HCI, the term anthropomorphism has also been used to describe the humanlike appearance or characteristics of a computer agent (e.g. Nowak and Rauh 2008). This means that an agent can behave, look or sound more or less anthropomorphic (i.e. humanlike). In HCI, views about the importance of anthropomorphism are not congruent. Some researchers suggest that the humanness of a computer makes interaction smoother and more motivating (Dehn and van Mulken 2000), whereas other researchers contend that humanlike features deceive users and make computer usage distressing (Shneiderman and Maes 1997). The findings of previous studies have also been contradictory, and they have mainly concentrated on studying the effects of agents' exterior features. Previous studies have investigated subjective experiences of agents' likeability, credibility and attractiveness when varying the level of anthropomorphism in visual appearance. Some of these studies have found that less anthropomorphic facial images are perceived as more credible and likeable than more anthropomorphic facial images, while some other studies have found that people perceive more anthropomorphic avatars as more attractive and credible than less anthropomorphic avatars (e.g. Nowak 2004, Nowak and Rauh 2005, 2008).

In light of the above, the role and significance of anthropomorphism in HCI is not clear, and it seems that the topic has to be approached with more fine-grained questions. One question that clearly needs to be studied more deeply is how the agent's level of anthropomorphism affects human emotional experiences. Weyers *et al.* (2006) found that avatars' dynamic (and therefore more realistic) happy facial expressions evoked stronger congruent facial reactions in participants than static expressions. Speech can also be synthesised so that the output sounds more or less humanlike, as described earlier. To the best of our knowledge, only two previous studies that have investigated the effects of the naturalness of speech synthesis on human emotions exist. The results of these studies showed that only more humanlike voice evoked emotion-specific facial muscle (Ilves and Surakka 2004) and pupil responses (Ilves and Surakka 2009) when the participants listened to sentences with negative, neutral and positive emotional content produced by two different speech synthesisers. In these studies, the sentences were produced by the synthesised voices without prosodic variation, so the emotions were expressed only by the emotional meaning of the

sentences. Further, because the two synthesisers did not differ with regard to prosodic features, it was possible to study the effects of the naturalness of the voice. The synthesis techniques used were diphone synthesis and synthesis based on the microphonemic method, which concatenates about 10 millisecond-long samples from natural speech. In addition, Ilves *et al.* (2011) have studied how increasing prosodic variation (and thus increasing the naturalness of speech) affects emotional experiences. The speech synthesiser they used was based on the unit selection technique. They found that emotional sentences spoken in a neutral speaking style evoked more emotion-relevant ratings of arousal than the sentences spoken in a monotonous speaking style. However, probably because the voice itself was very natural, these two speaking styles had quite small effects on emotional responses. For example, the ratings of valence were not affected by the speaking style. In order to partially replicate the previous findings and to extend the understanding of the effects of the naturalness of speech synthesis, the present study included all contemporary main speech synthesis techniques and used types of stimuli that differed from those used in previous studies. HMM-based synthesis system was not included in the present study, because of poor sound quality. After removing the prosodic variation from the HMM-based voice, buzz noise caused by a vocoder was invasive.

In addition to emotional pictures, facial images, sounds and brief texts, affective words have been used in evoking emotions in people. One of the most widely used affective word stimulus sets is the Affective Norms for English Words (ANEW) (Bradley and Lang 1999). In the study of Larsen *et al.* (2003), the ANEW words were visually presented, and they caused facial muscle reactions congruent with the emotional content of the stimuli. Buchanan *et al.* (2006) found that visually presented affective words evoked autonomic responses similar to those previously reported for emotional pictures and sounds. In some studies, the affective words were presented auditorily (e.g. Wambacq *et al.* 2004), but in these studies, the words were generated by a human speaker by varying the prosody of the voice. However, studies of the effects of emotional words presented by synthesised speech have not been reported.

Two prevalent theoretical approaches to categorising emotions exist. The discrete emotions theory posits the existence of separate, distinguishable emotional states, such as anger, fear, sadness and joy (Ekman 1992). The dimensional theory, on the other hand, argues that emotions can be defined through a certain set of dimensions, like valence and arousal (Bradley and Lang 1994). The valence dimension refers to the pleasantness of an emotional state, ranging from very negative to very positive. The arousal dimension refers to the arousal state, ranging from calm to aroused or excited. These two theoretical views are not necessarily mutually exclusive, since the discrete emotions can be located in a specific place in a two-dimensional affective space (Christie and Friedman 2004).

Based on the dimensional theory of emotion, a set of bipolar scales was formed (Bradley and Lang 1994). Using these scales, people can assess their subjective emotional experiences on a scale ranging from one to nine. The most frequently used scales are valence and arousal. On the valence scale, 1 denotes unpleasant experiences, while in the arousal scale, it denotes calming experiences. On the valence scale, 9 represents pleasant experiences, while on the arousal scale, it represents arousing experiences. It has been suggested that the valence and arousal dimensions are related to the motivational system of approach or avoidance (Lang *et al.* 1992). Although change in approach-avoidance tendency is one of the most important parts of emotional processing (see e.g. Mauss and Robinson 2009), previous studies have rarely collected the ratings of these motivational tendencies. The motivational tendency can be measured by collecting approach-withdrawal tendency ratings with a scale ranging from avoidable to approachable (e.g. Anttonen and Surakka 2005).

The aims of this study were as follows. First, it aimed to extend and deepen knowledge of emotional reactions to synthesised spoken messages, investigating whether it is possible to affect the subjective ratings of emotional experiences using the lexical emotional content of single words selected from a standardised stimulus set of affective words.[1] Second, it aimed to study how the naturalness of voice affects the ratings of emotions and voice quality. In contrast to earlier studies, the present study included all contemporary main speech synthesis techniques. Words with emotionally negative, neutral and positive content were created using three different speech synthesis techniques: formant synthesis, diphone synthesis and unit selection synthesis. The words were produced so that the voices were free from prosodic variation, resulting in a monotone tone of voice. Thus, it was possible to study whether the lexical content alone is sufficient to affect the emotion-related ratings and to determine whether the synthesisers or, in other words, the different levels of naturalness would affect the ratings of emotions. The experiment was a within subject repeated measures design.

2. Methods

2.1. Participants

Twenty-four volunteer students from a local university participated in the study (12 females and 12 males,

average age 25.2 (standard deviation 8.9), age range 18–51 years). The participants were recruited from an introductory computer science course and an introductory HCI course. Each participant was a native speaker of Finnish and had normal hearing (according to self-reports).

2.2. Equipment

The presentation and the rating of the stimulus were controlled by E-Prime© (Schneider *et al*. 2002) experiment generator software running on a PC computer with a Windows XP operating system. The stimuli were presented via two loudspeakers placed on the table in front of the participant.

The speech synthesisers were Finnish-speaking synthesisers called Ydinpuhe (Saarni 2010), Suopuhe (http://www.ling.helsinki.fi/suopuhe/english.shtml, Accessed 3 June 2011) and Mika (http://www.bitlips.fi/index.en.html, Accessed 3 June 2011). Ydinpuhe is a rule-based formant synthesiser, while Suopuhe and Mika are concatenative synthesisers. More precisely, Suopuhe is a diphone-based synthesis system, and Mika is based on the unit selection technique.

2.3. Stimuli

Stimulus words were carefully selected from the Affective Norms for English Words (ANEW) (Bradley and Lang 1999), which provides reference mean rating values for valence (positive/negative), arousal (high/low) and dominance (high/low) using nine-point bipolar scales. Fifteen stimulus words were selected and translated into Finnish (Table 1) so that five stimuli were strongly positive words (average valence = 8.3), five were strongly negative words (average valence = 1.9) and five were neutral words (average valence = 5.1). The average arousal rating for both the positive and the negative words was 7.0, and for the neutral words, it was 4.5. The words in the different emotion categories were matched according to word length.[2,3] The lengths of the words in each category were five, six (two words), seven and ten letters.

The words were produced by the male voices of the speech synthesisers. The different synthesisers produced the same words at the same rate. The length of a stimulus was 0.66 s on average, ranging between 0.47 and 0.94 s. The fundamental frequency (F0) of each synthesiser's voice was set to 100 Hz, and the volume was normalised at 75 dB using the Wave Surfer 1.8.5© program (Sjölander and Beskow 2000). The words were produced so that the F0 variation was set as flat as possible in order to keep the prosody of the voices as neutral as possible.

2.4. Experimental procedure

After being welcomed to the sound-attenuated laboratory, the participant was seated in a chair and informed that the aim of the experiment was to measure reactions to auditory stimuli. The participant was also told that the experiment consisted of two phases: a listening phase and a rating phase. The distance from the participant's eyes to the centre of the screen was adjusted to 90 cm.

In order to familiarise the participant with the monotone voice of synthetic speech, the Mikropuhe© (http://www.mikropuhe.com/, Accessed 3 June 2011) speech synthesiser was used to explain the experimental task to the participant. This synthesiser was not used in the actual experiment. The participant's task was to relax and carefully listen to a set of words generated by speech synthesis. In addition, the participant was instructed to indicate whether she/he understood the word she/he had just heard by pressing one of two buttons on a response box.

The listening phase proceeded as follows. A fixation-cross appeared at the centre of the screen 10 s before the onset of the first stimulus. During the rest of the experiment, the duration of the fixation-cross before the stimulus was five seconds. Five seconds from the stimulus offset, the fixation-cross disappeared, and the question 'Did you understand the word?' appeared on the screen. This indicated that the participant should press the left button of the response box if she/he did not understand the word and the right

Table 1. Words used in the experiment.

Word meaning					
Negative		Neutral		Positive	
Finnish	English	Finnish	English	Finnish	English
Pommi	Bomb	Hella	Stove	Nauru	Laughter
Pettää	Betray	Sakset	Scissors	Juhlia	Party
Hukkua	Drown	Tankki	Tank	Hauska	Fun
Pelokas	Afraid	Työkalu	Tool	Voittaa	Win
Terroristi	Terrorist	Teollisuus	Industry	Onnellinen	Lucky

button if she/he understood the word. Five seconds later, the participant heard a beep, and the fixation-cross appeared again to signal the end of the answering and resting period. All stimuli were presented randomly, and the listening phase of the experiment lasted approximately 13 min.

After the listening phase, the participant rated the stimuli. First, each participant rated her/his emotional experiences evoked by the stimuli on three dimensions: valence, arousal and approachability (see Appendix). All the stimuli heard once already in the listening phase were presented again (randomly), and the ratings were given on three nine-point bipolar scales. The valence scale ranged from a negative experience to a positive experience, the arousal scale from a calm experience to an aroused experience and the approach/withdrawal scale from a withdrawal experience to an approach experience. The centre of each scale represented a neutral point (e.g. neither unpleasant nor pleasant). First, the participant heard a word and gave a valence rating. Then, she/he heard the word again and gave an arousal rating. After the third repetition, the participant gave an approachability rating.

Second, the participant rated how pleasant, natural and clear the voice was. These questions were selected from a modified mean opinion score (MOS) scale (Viswanathan and Viswanathan 2005). The ratings were given on five-point scales to all the neutral words presented in the experiment.[4]

Before both rating sessions, two stimuli were rated to practice the rating process in order to ensure that the participant understood the idea behind the scales. For both ratings, the scales were presented on the computer screen, and a keyboard was used to give the ratings. After the stimulus ratings, the participant was debriefed about the purpose of the study, and she/he gave a written consent.

2.5. Data analysis

The subjective rating data were analysed using within-subject repeated measures ANOVAs. Greenhouse-Geisser adjusted degrees of freedom (rounded to the nearest whole number) were used when violations of sphericity occurred. For the multiple post-hoc comparisons, Bonferroni corrected pairwise *t*-tests were used. All data were analysed using SPSS® statistical software, version 17.0 (SPSS Inc., Chicago, IL).

3. Results

3.1. The number of the words that were not understood

During the listening phase, the participants answered the question of whether or not they had made out the word they had just heard. On average, 20.83% of the words were not understood. The averaged percentages and standard error of the mean (SEM) of the words that were not understood are presented in Figure 1. To test the differences between the synthesisers, a Friedman test was conducted. The Friedman test showed the statistically significant effect of synthesiser, $\chi^2(2) = 33.23$, $p < 0.001$. The Wilcoxon signed-rank test was used for pairwise comparisons. The test showed that there were significantly more words that were not understood when the words were produced via formant synthesis as compared to diphone synthesis, $Z = 4.06$, $p < 0.001$, and unit selection synthesis, $Z = 4.22$, $p < 0.001$. The difference between diphone synthesis and unit selection synthesis was not statistically significant $Z = 1.17$, ns.

Pairwise comparisons showed that the mean ratings for valence, $t(23) = 1.60$, NS, or arousal, $t(23) = 1.71$, NS, did not differ significantly between the words that were understood and the words that were not understood in the listening phase. Thus, the words that were judged as difficult to understand in the listening phase were not removed from the further analysis.

3.2. Ratings of emotional experiences

The mean ratings for valence are presented in Figure 2. For the ratings of valence, a 3 × 3 two-way (synthesiser × emotion category) ANOVA revealed a statistically significant main effect for emotion category, $F(1, 25) = 27.77$, $p < 0.001$, and a statistically significant interaction of synthesiser and category, $F(4, 92) = 6.19$, $p < 0.001$. The main effect of synthesiser was not statistically significant, $F(1, 28) = 2.84$, NS.

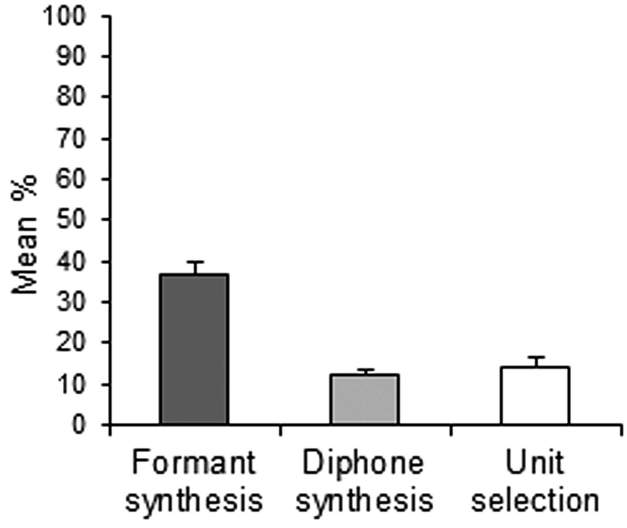

Figure 1. Average percentages (and S.E.M) of the words that were not understood.

As the interaction of synthesiser and emotion category was significant, one-way ANOVAs with emotion category as a factor were conducted separately for each voice. The effect of emotion category was statistically significant for all voices. Post-hoc pairwise comparisons (see Table 2) showed that the negative words were experienced as significantly more negative than the neutral and the positive words. Further, the positive words were experienced as significantly more positive than the neutral words.

Because of the significant interaction, three one-way ANOVAs with synthesiser as a factor were also conducted separately for the different emotion categories. The results showed that the effect of the synthesiser was statistically significant for the negative words, $F(2, 46) = 6.25$, $p = 0.004$. Post-hoc pairwise comparisons showed that the negative words produced via unit selection synthesis were rated as significantly more negative than the words produced via formant synthesis, $t(23) = 2.95$, $p = 0.022$. The differences between unit selection synthesis and diphone synthesis, $t(23) = 1.94$, NS, and between diphone synthesis and formant synthesis were not statistically significant, $t(23) = 2.11$, NS. The one-way ANOVA for the ratings of the neutral words was not statistically significant, $F(2, 46) = 2.24$, NS. For the ratings of the positive words, the effect of the synthesiser was statistically significant, $F(2, 46) = 3.24$, $p = 0.048$. Post-hoc pairwise comparisons showed that the positive words generated via the diphone synthesis were experienced as significantly more positive than the positive words generated via the formant synthesis, $t(23) = 3.28$, $p = 0.010$. The differences between diphone synthesis and unit selection synthesis, $t(23) = 0.64$, NS, and between unit selection synthesis and formant synthesis, $t(23) = 1.69$, NS, were not statistically significant.

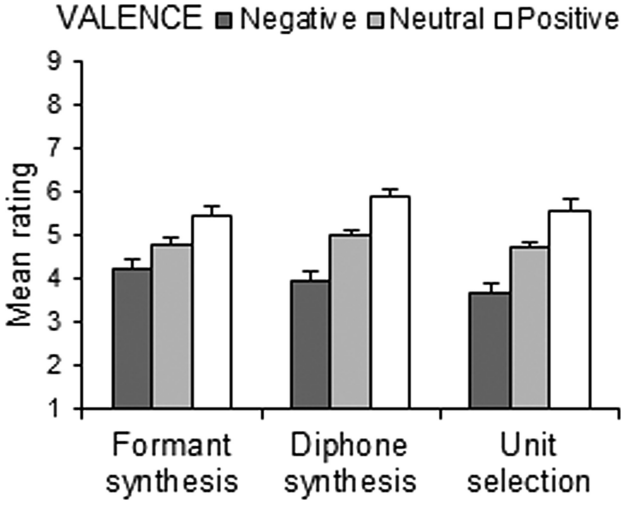

Figure 2. Mean ratings (and S.E.M) of valence.

The mean ratings of arousal are shown in Figure 3. For the ratings of arousal, a 3 × 3 two-way (synthesiser × emotion category) ANOVA showed a significant main effect for emotion category, $F(2, 46) = 12.89$, $p < 0.001$. The main effect for synthesiser, $F(2, 46) = 1.71$, NS, and the interaction of synthesiser and category, $F(4, 92) = 1.56$, NS, were not statistically significant. Post-hoc pairwise comparisons between the emotion categories showed that the negative words were rated as significantly more arousing than the neutral, $t(23) = 4.93$, $p < 0.001$, and the positive words, $t(23) = 4.06$, $p = 0.001$. The difference between the positive and the neutral words was not statistically significant, $t(23) = 0.02$, NS.

The average ratings of approachability are illustrated in Figure 4. A 3 × 3 two-way (synthesiser × emotion category) ANOVA revealed a statistically significant main effect for emotion category, $F(1, 30) = 26.38$, $p < 0.001$, and a statistically significant interaction for synthesiser and emotion category, $F(4, 92) = 4.04$, $p = 0.005$. The main effect for synthesiser was not statistically significant, $F(2, 35) = 3.45$, NS.

Due to the significant interaction of synthesiser and emotion category, one-way ANOVAs with emotion category as a factor were conducted separately for each voice. The effect of emotion category was statistically significant for all voices (see Table 3).

Table 2. Results of the effect of the synthesiser on the ratings of valence.

Synthesiser	ANOVA	Pairwise comparisons
Formant	$F(1, 29) = 16.03$, $p < 0.001$	Negative < Neutral $t(23) = 3.27$, $p = 0.010$ Negative < Positive $t(23) = 4.25$, $p < 0.001$ Neutral < Positive $t(23) = 3.97$, $p < 0.001$
Diphone	$F(1, 29) = 29.42$, $p < 0.001$	Negative < Neutral $t(23) = 5.20$, $p < 0.001$ Negative < Positive $t(23) = 5.75$, $p < 0.001$ Neutral < Positive $t(23) = 4.57$, $p < 0.001$
Unit selection	$F(1, 27) = 24.70$, $p < 0.001$	Negative < Neutral $t(23) = 5.97$, $p < 0.001$ Negative < Positive $t(23) = 5.20$, $p < 0.001$ Neutral < Positive $t(23) = 3.50$, $p = 0.006$

Because of the significant interaction, one-way ANOVAs with synthesiser as a factor were also conducted separately for the different emotion categories. The effect of synthesiser was significant for the ratings of the positive words, $F(2, 46) = 6.47$, $p = 0.003$. Post-hoc pairwise comparisons showed that the positive words generated via diphone synthesis were rated as significantly more approachable than the positive words generated via formant synthesis, $t(23) = 4.06$, $p = 0.001$. There were no statistically significant differences between diphone synthesis and unit selection synthesis, $t(23) = 1.79$, NS, or between unit selection synthesis and formant synthesis, $t(23) = 1.67$, NS. For the ratings of the neutral, $F(2, 46) = 2.79$, NS, and negative words, $F(2, 36) = 0.43$, NS, the effect of synthesiser was not statistically significant.

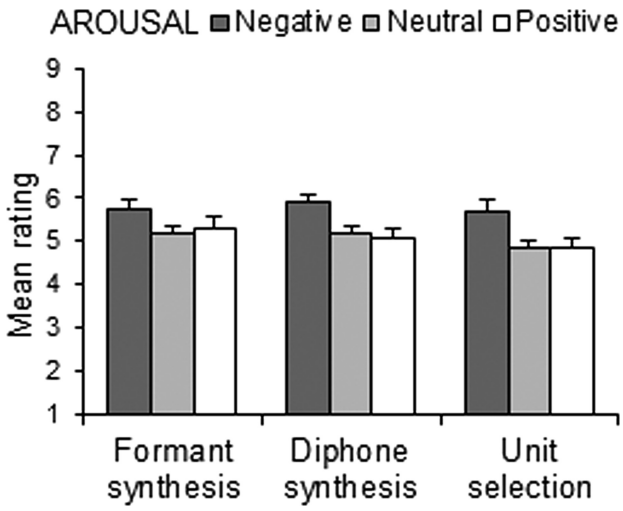

Figure 3. Mean ratings (and S.E.M) of arousal.

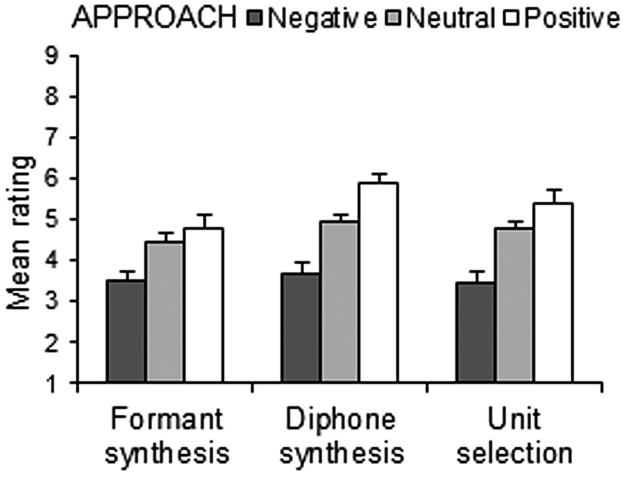

Figure 4. Mean ratings (and S.E.M) of approachability.

3.3. *Subjective ratings of the voice*

The average ratings of the pleasantness, naturalness and clarity of the voice are shown in Figure 5. Three one-way ANOVAs with speech synthesiser as a factor were performed separately for the ratings of the pleasantness, naturalness and clarity of the voice. For the ratings of pleasantness there was a statistically significant effect of synthesiser, $F(2, 46) = 16.8$, $p < 0.001$. Post-hoc pairwise comparisons showed that the voice of formant synthesis was experienced as significantly less pleasant than the voice of diphone synthesis, $t(23) = 5.12$, $p < 0.001$, and the voice of unit selection synthesis, $t(23) = 4.45$, $p < 0.001$. The difference between diphone synthesis and unit selection synthesis was not statistically significant, $t(23) = 0.43$, NS.

The one-way ANOVA for the ratings of naturalness revealed a statistically significant effect of synthesiser, $F(2, 46) = 97.93$, $p < 0.001$. Post-hoc pairwise comparisons showed that the voice of unit selection synthesis was rated as significantly more natural than the voice of formant synthesis, $t(23) = 12.15$, $p < 0.001$, and the voice of diphone synthesis, $t(23) = 3.28$, $p = 0.01$. The voice of diphone synthesis was rated as significantly more natural than the voice of formant synthesis, $t(23) = 11.11$, $p < 0.001$.

Table 3. Results of the effect of the synthesiser on the ratings of approachability.

Synthesiser	ANOVA	Pairwise comparisons
Formant	$F(1, 33) = 13.02$, $p < 0.001$	Negative < Neutral $t(23) = 5.77$, $p < 0.001$ Negative < Positive $t(23) = 4.05$, $p = 0.002$ Neutral vs. Positive $t(23) = 1.21$, ns
Diphone	$F(1, 32) = 30.88$, $p < 0.001$	Negative < Neutral $t(23) = 5.75$, $p < 0.001$ Negative < Positive $t(23) = 6.10$, $p < 0.001$ Neutral < Positive $t(23) = 3.82$, $p = 0.003$
Unit selection	$F(2, 36) = 18.27$, $p < 0.001$	Negative < Neutral $t(23) = 5.14$, $p < 0.001$ Negative < Positive $t(23) = 4.81$, $p < 0.001$ Neutral vs. Positive $t(23) = 2.04$, ns

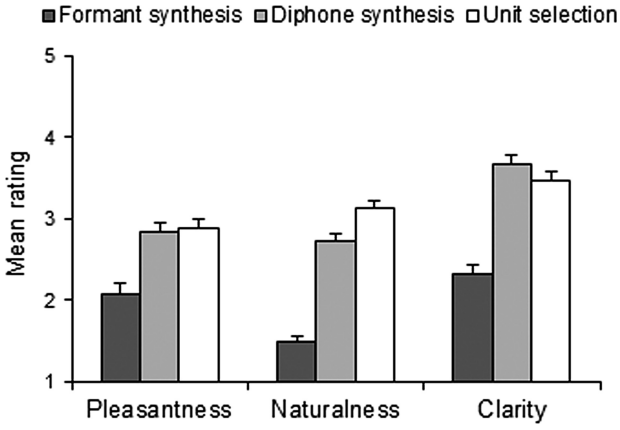

Figure 5. Mean ratings (and S.E.M) of pleasantness, naturalness and clarity of voice for all three speech synthesisers.

Moreover, for the ratings of the clarity of the voice, the effect of synthesiser was statistically significant, $F(2, 46) = 67.20$, $p < 0.001$. Post-hoc pairwise comparisons showed that the voice of formant synthesis was rated as significantly less clear than the voice of diphone synthesis, $t(23) = 10.34$, $p < 0.001$, and the voice of unit selection synthesis, $t(23) = 8.64$, $p < 0.001$. The difference between diphone synthesis and unit selection synthesis was not statistically significant, $t(23) = 1.84$, NS.

4. Discussion

The results showed that the affective words produced by the synthesised speech evoked significantly different ratings of emotional valence, arousal and approachability. The words with negative emotional content resulted in more negative ratings of valence and lower ratings of approachability than the words with neutral or positive content. The words with positive emotional content resulted in more positive ratings of valence than the words with neutral content. In addition, the words with negative emotional content evoked higher ratings of arousal than the words with neutral or positive content. These findings were true for all synthesisers. Overall, the negative words evoked both higher ratings of negative experiences and higher ratings of withdrawal from the stimuli. The positive words evoked both more positively valenced experience ratings and higher ratings of approach. Thus, there was a noticeable connection between valence ratings and approachability ratings. As mentioned in the introduction, it has been suggested that the valence and arousal dimensions are related to the motivational system of approach or avoidance (Lang *et al.* 1992). In the present experiment, this motivational tendency was measured by the ratings of the approach-withdrawal tendency using a scale ranging from avoidable to approachable. The connection between valence and approachability is not always straightforward. For example, anger can be seen as a negatively valenced but approach-related emotion (e.g. Harmon-Jones and Sigelman 2001). However, generally, the valence and approach motivation are related such that positive emotions are related to the approach tendency and negative emotions are related to the withdrawal tendency (Lang *et al.* 1992, Elliot 1999). This was also the case in the present study. Thus, the synthetically produced affective words modulated the ratings of valence and arousal as well as the ratings of motivational tendency to approach or withdraw from the stimuli. In many ways, these findings were expected and confirm the results of the previous studies (e.g. Ilves and Surakka 2004, 2009, Ilves *et al.* 2011).

An even more important finding is that the synthesis techniques used affected the ratings of emotions, such that the emotionally negative and positive words produced by the concatenative synthesisers, the more humanlike synthesisers, resulted in more pronounced ratings of emotions than the words produced by the formant synthesiser. Specifically, experienced valence and approachability were rated as significantly higher after the positive words produced via diphone synthesis than after the positive words produced via formant synthesis. In addition, experienced valence was rated as significantly lower after the negative words produced via unit selection synthesis than after the negative words produced via formant synthesis. Therefore, the voice had an influence on both the ratings of the words with positive emotional content and the ratings of the words with negative emotional content. It could be argued that the ratings of valence were affected by the difficulty involved in understanding the words produced by the different synthesisers. However, there were no statistically significant differences in the ratings of valence and arousal between the words that were rated as clear and those that were rated as unclear.

The ratings of the voices revealed that the unit selection synthesis voice was assessed as the most natural. Further, the formant synthesis voice was rated as the most unnatural. Thus, even though the prosodic cues were removed from the voices, the unit selection synthesis voice was assessed as significantly more natural than the voices of the other synthesisers. These ratings confirmed that the naturalness of the voices was experienced as expected. There were also significant differences between the synthesisers concerning the intelligibility and the ratings of the pleasantness and clarity of the voice. The voices of both concatenative synthesisers were rated as significantly more

understandable, pleasant and clear than the voice of the formant synthesiser.

The results of this study are in line with previous findings that the ratings of emotional experiences can be affected by the lexical content of verbal stimuli. In the study of Buchanan *et al.* (2006), the emotional category of visually presented words affected the ratings of valence and arousal as expected. In addition, their results showed that autonomic responses to the visually presented emotional words were similar to those previously reported for emotional pictures and sounds. Larsen *et al.* (2003) found that the experienced valence of visually presented words correlated linearly with the activity of the facial muscle associated with frowning. This study also supports previous findings (Ilves and Surakka 2004, 2009, Ilves *et al.* 2011) that synthesised speech with emotional content can modulate the ratings of emotions. In this study, the standardised set of short affective words was used instead of emotional sentences, so a single synthetically produced word alone was sufficient to affect at least the subjective ratings of emotional experiences. In addition, even though our experiment was a highly controlled laboratory study, the affective stimuli may also evoke similar responses in less-controlled environments. This is supported by the findings of Knoll *et al.* (2011), who found that the affective ratings of speech samples were similar for the samples rated in a laboratory and those rated in a more ecologically valid environment (i.e. the Internet).

The arousal ratings of the negative words were significantly higher than the arousal ratings of the neutral and positive words, although the reference arousal values in the ANEW were equal for the negative and positive words. Some comparable previous studies have also found that there is a tendency to rate negative stimuli as more arousing than positive stimuli. In the study of Ilves *et al.* (2011), negative sentences generated by a synthesised voice were experienced as more arousing than positive sentences. Bertels *et al.* (2009) found that negative words spoken by a human voice evoked higher ratings of arousal than positive words. There have also been similar findings for other types of stimuli, such as pictures with emotion-related content. Several studies have found that participants rate negative stimuli as more arousing than positive stimuli, though the stimuli were selected from a standardised stimulus set so that the mean arousal ratings of negative and positive stimuli matched each other (e.g. Canli *et al.* 1998, Anttonen and Surakka 2005, Grühn and Scheibe 2008). The reason for this is unknown, but it may relate to a contextual effect. After a small subset of stimuli is selected from a large stimulus set, the context of stimulus ratings is different from the original context; thus, changing the stimulus space for giving ratings in turn, changes the values for ratings.

In the present study, synthesis technique had an effect on the ratings of valence and approachability, but it did not affect the arousal ratings. Instead, in previous studies, the degree of naturalness has affected the ratings of arousal but not the ratings of valence (Ilves and Surakka 2009, Ilves *et al.* 2011). The present study and the previous studies differ in at least two respects. In the present study, the stimuli consisted of affective words, whereas in the previous studies, the stimuli were emotional sentences. Therefore, the different nature of the stimuli may have affected the ratings. In addition, all the voices in the previous studies have been based on some concatenative method and thus have been quite natural. The study by Ilves and Surakka (2009) utilised two concatenative synthesisers. In the study of Ilves *et al.* (2011), the naturalness of the voice was further increased by increasing the prosodic variation of the voice. Thus, in light of these findings, it seems that when the voice is sufficiently humanlike, further increasing the naturalness by using a technique that sounds even more humanlike or increasing the prosodic variation of the voice does not substantially enhance the experienced valence. Other findings also support this interpretation. Bertels *et al.* (2009) found that people rated the valence of positive words similarly despite whether the words were uttered by a neutral tone of human voice or an emotionally congruent tone of human voice. As mentioned, in the present study, the synthesis technique had no significant effect on arousal ratings. One reason for that may result from the fact that the words were not initially arousing enough.

Emotions are known to consist of three components, including physiological, behavioural and experiential changes (Frijda 1988). In the present study, the method selected to study emotions was the subjective rating of emotions. In other words, the task of the participants was to rate their emotional experiences evoked by the stimuli with the help of valence, arousal and approachability scales. This procedure is widely used and accepted when evaluating subjective emotional experience (Mauss and Robinson 2009). The ratings of emotions showed relatively clearly that the words with emotional content resulted in different ratings of emotional valence, arousal and approachability. More importantly, the synthesis technique affected the emotion-related ratings. The differences in the ratings of emotions may reflect the fact that there were changes in spontaneous emotional experience as well or it is possible that only the cognitive evaluation of the experience changed without the changes in spontaneous experience. Alternatively, it is possible that both spontaneous experience and

cognitive evaluation changed. The effect of synthesised speech with lexical emotional content on humans is an under-investigated subject. Based on the current results, we can say that the emotional messages produced by the speech synthesisers evoked the differences, at least in the cognitive ratings of emotions. However, the subjective rating of current emotion is always only one side of the emotion, and a more comprehensive understanding of spontaneous emotional experiences can be achieved by using multiple measurements, such as measuring physiological activation in addition to subjective feelings. On the other hand, it has been frequently found that the subjective ratings of emotions correlate with physiological changes (e.g. Lang *et al*. 1993, Larsen *et al*. 2003).

An additional note of caution is that we cannot be sure if the participants rated their own personal emotional reactions to the words as instructed or if the ratings related more to the ratings of the meanings of the words (e.g. how pleasant the word itself was rather than how pleasant the reaction the word evoked was). In any event, the results showed significant differences between the synthesisers, such that the ratings of experienced emotions were stronger when the speaking voice was more humanlike. These differences between the synthesisers were specifically related to the differences in the ratings of the negative and positive words. The synthesiser had no significant effect on the ratings of the neutral words (i.e. the neutral words were rated equally neutral regardless of the synthesis technique used).

Previous findings concerning anthropomorphism have been somewhat inconsistent. The current results showing that the more humanlike voices were rated as more natural and pleasant and that the more humanlike voices were associated with more pronounced ratings of emotions are in line with the findings that emphasise the importance of anthropomorphism. In some previous studies, it has been found, for example, that avatars that are more anthropomorphic are conceived as more attractive than less anthropomorphic avatars (Nowak and Rauh 2005, 2008). In some studies, the less anthropomorphic facial images were rated as more likeable than the more anthropomorphic facial images (Nowak 2004). One well known phenomenon related to the human likeness of a robot or virtual character and human emotional response is the uncanny valley phenomenon. The term derives from the hypothesis presented by Japanese roboticist Masahiro Mori (1970). The uncanny valley hypothesis suggests that the more a robot or a virtual character looks or acts like a real human the more positively people perceive it until a certain point when it causes a negative response among human observers. This valley-shaped dip in a proposed graph represents a point when the appearance of the robot or virtual character is almost but not fully humanlike. In our study, instead, the more anthropomorphic voices were rated as more pleasant and natural than the less anthropomorphic voice. Further, in previous studies, it has been found that more humanlike voice elicits more pronounced emotion-related ratings and physiological reactions in people (Ilves and Surakka 2004, 2009). Thus, even though the present results showed that even the purely artificially produced machinelike voice manages to modulate the ratings of emotional experiences, they also showed that by increasing the naturalness of speech, it is possible to enhance the effects of synthetic speech on emotional experiences.

Although the voice created via unit selection synthesis was assessed as more natural than the voice of the other concatenative synthesiser diphone synthesis, other differences between these synthesisers were not significant. The voices created via unit selection synthesis and diphone synthesis were rated as equally pleasant and clear, and there were no significant differences in the emotional ratings between these synthesisers. Therefore, in light of this and previous studies (Ilves and Surakka 2004, 2009), it seems that the speech produced via the diphone synthesis offers sufficiently natural speech to evoke positive feelings and emotions. This is noteworthy, because the advantage of diphone synthesis over unit selection synthesis is that diphone synthesis uses a much smaller database of speech recordings and, thus, has lower computational requirements than unit selection synthesis.

On the other hand, it is interesting to speculate as to why the differences in the ratings of the positive words were found specifically between diphone and formant synthesis whereas the differences in the ratings of the negative words were found between unit selection and formant synthesis. In other words, the effects of the positive words on the ratings of valence and approachability were enhanced when the words were produced via diphone synthesis, and the effects of the negative words on the ratings of valence were enhanced when the words were produced via the unit selection synthesis as compared to formant synthesis. Why did unit selection synthesis, the most natural voice, not also evoke the most positive ratings of the positive words? It has been found that without any affective cues, human speech is usually interpreted as unfriendly (Waaramaa-Mäki-Kulmala 2009, p. 74). Bertels *et al*. (2009) have also found that negative words uttered in a neutral tone of human voice are experienced as more threatening than negative words uttered in an emotionally congruent tone of human voice, possibly because the neutral tone of voice evokes an impression of coldness. Thus, it is possible that the

humanlike but flat voice produced via unit selection synthesis was experienced as harsh and thus enhanced the effects of the negative words, whereas the flat speaking style was more tolerable for the less humanlike voice, like diphone synthesis, as compared to the flat speaking style of the more humanlike voice. This finding also suggests that the relations between anthropomorphism and human behaviour are not necessarily straightforward.

The affective words in this study were produced by the flat, monotone voices as mentioned, and in spite of that, the positive and negative words evoked different emotional experience ratings. Although statistically significant, clear differences were found, the ratings of valence, arousal and approachability were concentrated quite near to the centre of the scale. This was quite surprising, because the positive and negative words were selected from the ANEW based on their high mean rating on the arousal scale and their high or low mean rating on the valence scale. For example, the mean valence was 8.25 for the selected positive words and 1.88 for the negative words, but in the present study, valence ratings were quite near neutral. Thus, in the future, it would be interesting to study emotional reactions to the emotional messages comparing the differences between the synthesised voice and the textual manner of representation. This kind of study would clarify whether the knowledge does speech enhance the emotional impact of a stimulus or whether synthesised speech flattens the emotional content of a stimulus, as seemed to be the case in the current study. However, one explanation why the present ratings of emotion were more neutral than the original ANEW ratings may relate to habituation effect. In the present experiment, the participants had already heard all words in the listening phase before they gave emotional ratings for them. There is evidence that physiological responses, such as skin conductivity or amygdala activation, habituate following repeated presentation of emotional stimuli (e.g. Bradley *et al*. 1993, Breiter *et al*. 1996). Thus, it is possible that the ratings of emotions would have been stronger if the ratings would have been given after the first presentation.

Another interesting future study would be a comparison of human and synthesised speech in order to understand more thoroughly the effects of voice naturalness on human emotions. On the other hand, one advantage of speech synthesisers over human speech is their controllability. Speech synthesisers offer a high level of control over prosodic cues such as speech rate, volume, voice quality and F0 range. It is likely that adding emotion-related prosodic cues would enhance the effects of synthesisers on emotional experiences. Thus, carefully manipulating the prosodic cues of synthetic voices could be one way to study the effects of voice naturalness further.

Overall, our results showed that synthesised speech with lexical emotional content affects emotion-related ratings. The study of human emotions, including emotional experiences, has a long history. The present results showed that by measuring emotional experiences, it was possible to reveal differences between different synthesis techniques. The method seems to be valuable when studying human experiences in HCI. On the other hand, our results showed that even the use of a monotone, machinelike voice can result in different emotion-related ratings. This may derive from the importance of speech in human communication. Speech is a great medium when people cooperate with others; people do not merely express themselves, but also influence others through speech. In addition, because hearing is not restricted by the position of the ears (i.e. as seeing is dependent on the position of the head and eyes), it is considered an important input channel for warning signals (e.g. Miller 1974, Scharf 1998). Emotions have served as a central mechanism for survival, and thus, people are likely highly responsive to emotional messages delivered through speech. It has been suggested that because of the intrinsic sensitivity to received spoken messages, people respond similarly whether the speech comes from a human or a machine (Nass and Brave 2005). The present evidence showed that synthetic spoken emotional cues effectively evoke emotional experiences. Thus, spoken messages from computers evoke both sociality (the CASA paradigm) and emotionality in their users.

5. Summary of the findings

The present results showed that the affective words produced by formant synthesis, diphone synthesis and unit selection synthesis evoked congruent emotion-related ratings in the participants. Thus, even the use of a monotone, machinelike voice can result in different emotion-related ratings. However, the ratings of valence and approachability were statistically significantly stronger when the affective words were generated by the more humanlike voices as compared to the more machinelike voice. Therefore, the voice features must be optimised to strengthen emotional communication between computers and humans.

Acknowledgements

The authors would like to thank all the participants of the study. This research was supported by the Doctoral Program in User-Centred Information Technology (UCIT) and a grant from the University of Tampere.

Notes

1. This study is a part of the larger study. In this article, the subjective ratings of the emotional words are presented.
2. The length of the words was counted as letters instead of phonemes because the pronunciation in Finnish language is very regular as compared to English language, for example. Generally, that means that one letter corresponds to one sound, that is, the Finnish words are pronounced as they are written.
3. The selected words consist of adjectives, verbs and nouns, because the selection of the words was based on the length of a word and the mean rating values for valence and arousal.
4. The reason why the speech quality ratings were given only for the neutral words was the duration of the experiment. The experiment took approximately 1 hour and 15 min. If the participants had given the speech quality ratings also for the negative and positive words it would have resulted in 90 more ratings causing too long and exhausting experiment.

References

Alais, D., Morrone, C., and Burr, D., 2006. Separate attentional resources for vision and audition. *Proceedings of the Royal Society*, 273 (1592), 1339–1345.

Anttonen, J. and Surakka, V., 2005. Emotions and heart rate while sitting on a chair. *In*: *Proceedings of the CHI 2005 SIGCHI conference on human factors in computing systems*, 2–7 April. New York: ACM, 491–499.

Aula, A. and Surakka, V., 2002. Auditory emotional feedback facilitates human–computer interaction. *In*: *Proceedings of the HCI 2002 international conference on human–computer interaction*, 2–6 September. London: Springer-Verlag, 337–349.

Bartneck, C., 2001. How convincing is Mr Data's smile: affective expressions of machines. *User Modeling and User-Adapted Interaction*, 11 (4), 279–295.

Beale, R. and Creed, C., 2009. Affective interaction: how emotional agents affect users. *International Journal of Human-Computer Studies*, 67 (9), 755–776.

Bertels, J., Kolinsky, R., and Morais, J., 2009. Norms of emotional valence, arousal, threat value and shock value for 80 spoken French words: comparison between neutral and emotional tones of voice. *Psychologica Belgica*, 49 (1), 19–40.

Bradley, M.M. and Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25 (1), 49–59.

Bradley, M.M. and Lang, P.J., 1999. Affective norms for English words (ANEW): stimuli, instruction manual and affective ratings. *Technical report C-1*. Gainesville: University of Florida.

Bradley, M.M., Lang, P.J., and Cuthbert, B.N., 1993. Emotion, novelty, and the startle reflex: habituation in humans. *Behavioral Neuroscience*, 107 (6), 970–980.

Brave, S., Nass, C., and Hutchinson, K., 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human–Computer Studies*, 62 (2), 161–178.

Breiter, H.C., *et al*., 1996. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, 17 (5), 875–887.

Buchanan, T.W., *et al*., 2006. The influence of autonomic arousal and semantic relatedness on memory for emotional words. *International Journal of Psychophysiology*, 61 (1), 26–33.

Cahn, J., 1990. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8 (1), 1–19.

Calvo, R.A. and D'Mello, S., 2010. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1 (1), 18–37.

Canli, T., *et al*., 1998. Hemispheric asymmetry for emotional stimuli detected with fMRI. *Neuroreport*, 9 (14), 3233–3239.

Christie, I.C. and Friedman, B.H., 2004. Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *International Journal of Psychophysiology*, 51 (2), 143–153.

Conati, C. and Maclaren, H., 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19 (3), 267–303.

Cowie, R. and Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40 (1–2), 5–32.

Dai, L., *et al*., 2005. Speech-based cursor control using grids: modeling performance and comparisons with other solutions. *Behaviour & Information Technology*, 24 (3), 219–230.

Dehn, D.M. and van Mulken, S., 2000. The impact of animated interface agents: a review of empirical research. *International Journal of Human–Computer Studies*, 52 (1), 1–22.

Dutoit, T., 1997. *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer Academic Publisher.

Ekman, P., 1992. An argument for basic emotions. *Cognition and Emotion*, 6 (3/4), 169–200.

Elliot, A., 1999. Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34 (3), 169–189.

Estrada, A.E., Isen, A.M., and Young, M.J., 1994. Positive affect improves creative problem solving and influences reported source of practice satisfaction in physicians. *Motivation and Emotion*, 18 (4), 285–299.

Fredrickson, B.L. and Branigan, C., 2005. Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion*, 19 (3), 313–332.

Frijda, N.H., 1988. The laws of emotion. *American Psychologist*, 43 (5), 349–358.

Grühn, D. and Scheibe, S., 2008. Age-related differences in valence and arousal ratings of pictures from the International Affective Picture System (IAPS): do ratings become more extreme with age? *Behavior Research Methods*, 40 (2), 512–521.

Guthrie, S.E., 1993. *Faces in the clouds: a new theory of religion*. New York: Oxford University Press.

Hassenzahl, M. and Tractinsky, N., 2006. User experience – a research agenda. *Behavior & Information Technology*, 25 (2), 91–97.

Harmon-Jones, E. and Siegelman, J., 2001. State anger and prefrontal brain activity: evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression. *Journal of Personality and Social Psychology*, 80 (5), 797–803.

Iida, A., *et al*., 2003. A corpus-based speech synthesis system with emotion. *Speech Communication*, 40 (1–2), 161–187.

Ilves, M. and Surakka, V., 2004. Subjective and physiological responses to emotional content of synthesized speech. *In*: N. Magnenat-Thalmann, C. Joslin, and H. Kim, eds. *Proceedings of the CASA 2004 international conference on computer animation and social agents*, 7–9 July. Geneva: Computer Graphics Society, 19–26.

Ilves, M. and Surakka, V., 2009. Emotions, anthropomorphism of speech synthesis, and psychophysiology. *In*: K. Izdebski, ed. *Emotions in the human voice. Volume III: culture and perception*. San Diego: Plural Publishing Inc., 137–152.

Ilves, M., Surakka, V., and Vanhala, T., 2011. The effects of emotionally worded synthesized speech on the ratings of emotions and voice quality. *In*: S. D'Mello *et al*., eds. *ACII 2011, Part I, lecture notes in computer science*, 6974. Berlin Heidelberg: Springer-Verlag, 588–598.

Isen, A.M., Daubman, K.A., and Nowicki, G.P., 1987. Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52 (6), 1122–1131.

Kappas, A., Hess, U., and Scherer, K.R., 1991. Voice and emotion. *In*: R.S. Feldman and B. Rimé, eds. *Fundamentals of nonverbal behavior. Studies in emotion & social interaction*. Cambridge: Cambridge University Press, 201–238.

Klein, J., Moon, Y., and Picard, R.W., 2002. This computer responds to user frustration: theory, design, and results. *Interacting with Computers*, 14 (2), 119–140.

Knoll, M.A., Uther, M., and Costall, A., 2011. Using the Internet for speech research: an evaluative study examining affect in speech. *Behaviour and Information Technology*. [online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/0144929X.2011.577192 [Accessed 29 August 2011].

Ku, J., *et al*., 2005. Experimental results of affective valence and arousal to avatar's facial expressions. *Cyberpsychology & Behavior*, 8 (5), 493–503.

Laird, J.D., *et al*., 1982. Remembering what you feel: effects of emotion on memory. *Journal of Personality and Social Psychology*, 42 (4), 646–657.

Lang, P.J., Bradley, M.M., and Cuthbert, B.N., 1992. A motivational analysis of emotion: reflex-cortex connections. *Psychological Science*, 3 (1), 44–49.

Lang, P.J., *et al*., 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30 (3), 261–273.

Larsen, J.T., Norris, C.J., and Cacioppo, J.T., 2003. Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, 40 (5), 776–785.

Law, E.L.-C. and van Schaik, P., 2010. Modeling user experience – An agenda for research and practice. *Interacting with Computers*, 22 (5), 313–322.

Mauss, I.B. and Robinson, M.D., 2009. Measures of emotion: a review. *Cognition and Emotion*, 23 (2), 209–237.

Miller, J.D., 1974. Effects of noise on people. *Journal of the Acoustical Society of America*, 56 (3), 729–764.

Minnema, M.T. and Knowlton, B.J., 2008. Directed forgetting of emotional words. *Emotion*, 8 (5), 643–652.

Mori, M., 1970. Bukimi no tani (The uncanny valley). *Energy*, 7 (4), 33–35. (MacDorman, K.F. and Minato, T., Translated).

Murray, I.R. and Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of Acoustical Society of America*, 93 (2), 1097–1108.

Murray, I.R. and Arnott, J.L., 1995. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16 (4), 369–390.

Nass, C. and Brave, S., 2005. *Wired for speech: how voice activates and advances the human–computer relationship*. Cambridge: The MIT Press.

Nass, C. and Lee, K.M., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7 (3), 171–181.

Nass, C., Steuer, J., and Tauber, E.R., 1994. Computers are social actors. *In*: *Proceedings of the CHI 1994 SIGCHI conference on human factors in computing systems*, 24–28 April. New York: ACM, 72–78.

Nowak, K.L., 2004. The influence of anthropomorphism and agency on social judgment in virtual environments. *Journal of Computer-Mediated Communication [online]*9. Available from: http://www3.interscience.wiley.com/cgi-bin/fulltext/120837918/HTMLSTART [Accessed 3 June 2011].

Nowak, K.L. and Rauh, C., 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication*, 11 (1), 153–178.

Nowak, K.L. and Rauh, C., 2008. Choose your "buddy icon" carefully: the influence of avatar androgyny, anthropomorphism and credibility in online interactions. *Computers in Human Behavior*, 24 (4), 1473–1493.

Partala, T. and Surakka, V., 2004. The effects of affective interventions in human-computer interaction. *Interacting with Computers*, 16 (2), 295–309.

Picard, R.W., 1995. *Affective computing*. Massachusetts: The MIT Press.

Prendinger, H., Mori, J., and Ishizuka, M., 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International Journal of Human–Computer Studies*, 62 (2), 231–245.

Reeves, B. and Nass, C., 1996. *The media equation: how people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.

Saarni, T., 2010. *Segmental durations of speech*. Thesis (PhD). University of Turku. Available from: https://www.doria.fi/handle/10024/52552 [Accessed 11 January 2012].

Scharf, B., 1998. Auditory attention: the psychoacoustical approach. *In*: H. Pashler, ed. *Attention*. Hove: Psychology Press, 75–117.

Schneider, W., Eschman, A., and Zuccolotto, A., 2002. *E-prime user's guide*. Pittsburgh: Psychology software Tools Inc.

Schröder, M., 2009. Approaches to emotional expressivity in synthetic speech. *In*: K. Izdebski, ed. *Emotions in the human voice. Volume III: culture and perception*. San Diego: Plural Publishing Inc., 307–321.

Shi, Y. and Zhou, L., 2009. Supporting dictation speech recognition error correction: the impact of external information. *Behaviour and Information Technology* [online]. Available from: http://www.tandfonline.com/doi/abs/10.1080/01449290903353039 [Accessed 30 August 2011].

Shneiderman, B. and Maes, P., 1997. Direct manipulation vs. interface agents. *Interactions*, 4 (6), 42–61.
Sjölander, K. and Beskow, J., 2000. Wavesurfer – an open source speech tool. *In*: *Proceedings of the ICSLP 2000 international conference on spoken language processing*, 16–20 October, Beijing, China, 464–467,
Spering, M., Wagener, D., and Funke, J., 2005. The role of emotions in complex problem-solving. *Cognition and Emotion*, 19 (8), 1252–1261.
Surakka, V. and Hietanen, J.K., 1998. Facial and emotional reactions to Duchenne and non-Duchenne smiles. *International Journal of Psychophysiology*, 29 (1), 23–33.
Viswanathan, M. and Viswanathan, M., 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language*, 19 (1), 55–83.
Waaramaa-Mäki-Kulmala, T., 2009. *Emotions in voice; Acoustic and perceptual analysis of voice quality in the vocal expression of emotions*. Thesis (PhD). University of Tampere. Available from: http://acta.uta.fi/pdf/978-951-44-7667-9.pdf [Accessed 1 September 2011].
Wambacq, I.J., Shea-Miller, K.J., and Abubakr, A., 2004. Non-voluntary and voluntary processing of emotional prosody: an event-related potentials study. *Cognitive Neuroscience and Neuropsychology*, 15 (3), 555–559.
Werner, N.S., Duschek, S., and Schandry, R., 2009. Relationship between affective states and decision-making. *International Journal of Psychophysiology*, 74 (3), 259–265.
Weyers, P., *et al*., 2006. Electromyographic responses to static and dynamic avatar emotional facial expressions. *Psychophysiology*, 43 (5), 450–453.
Wickens, C.D., 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3 (2), 159–177.
Zajonc, R.B., 1980. Feeling and thinking: preferences need no inferences. *American Psychologist*, 35 (2), 151–175.

Appendix

A short description of the instructions that were given to the participants before they gave the ratings of valence, arousal and approachability.

1. Valence

With this scale your task is to rate the valence of your experiences. The extremes of this scale are negative and positive. The negative end of the scale represents unpleasant experiences such as anger or fear. The positive end of the scale represents pleasant experiences such as enjoyment or happiness. You can select any number between one and nine depending on how strong your experience was. The bigger the number the more pleasant the experience is. The centre of the scale represents a point in which the experience was neither negative nor positive.

2. Arousal

With this scale, you will rate how much the word aroused you. The lower end of the scale represents the experiences of calmness, relaxation or sleepiness, for example. The higher end of the scale represents the experiences of excitement, frenzied or jittery, for example. As with the valence scale, you can select any number between one and nine depending on how calm or aroused you felt. The centre of the scale represents a point in which the experience was neither calm nor aroused.

3. Approachability

With this scale, you will rate a tendency to approach or avoid the word. The lower end of the scale represents a tendency to avoid listening of the word. The higher end of the scale represents a tendency to approach the word. The centre of the scale represents a point in which the experience of approach or avoid was neutral.