Mirja Ilves

Human Responses to Machine-Generated Speech with Emotional Content

ACADEMIC DISSERTATION
To be presented with the permission of the School of Information Sciences of the University of Tampere, for public discussion in the Pinni auditorium B1096 on June 19th, 2013, at noon.

School of Information Sciences
University of Tampere

Dissertations in Interactive Technology, Number 15
Tampere 2013

ACADEMIC DISSERTATION IN INTERACTIVE TECHNOLOGY

Supervisor:	Professor Veikko Surakka, Ph.D., School of Information Sciences, University of Tampere, Finland
Opponent:	Professor Emeritus Göte Nyman, Ph.D., Institute of Behavioural Sciences, Psychology University of Helsinki, Finland
Reviewers:	Professor Timo Saari, D.Soc.Sc., Department of Software Systems, Tampere University of Technology, Finland Docent Martti Vainio, Ph.D., Institute of Behavioural Sciences, Speech Sciences University of Helsinki, Finland

Dissertations in Interactive Technology, Number 15

School of Information Sciences
FI-33014 University of Tampere
FINLAND

ISBN 978-951-44-9149-8
ISSN 1795-9489

Suomen Yliopistopaino Oy – Juvenes Print
Tampere 2013

Abstract

The aim of the present thesis was to examine how people respond to synthetically produced lexical expressions of emotions. When speaking, both the content of spoken words and the prosodic cues, such as pitch and the speed of the speech, can mediate emotion-related information. To study how the pure content of spoken words affects human emotions, speech synthesizers offer good opportunities as they allow for good controllability over the prosodic cues. Synthetic speech can be generated using different techniques. Such speech can be purely machine generated or it can be based on different types (i.e. shorter or longer) of samples from human speech. On the basis of synthesis techniques, synthesizers can be classified according to the degree of human-likeness of the voice. Four different speech synthesizers were employed in this study, which all differed in their speech-production techniques. This also enabled an examination of the effects of the human-likeness of synthetic voices on human emotions.

Three key reasons motivated this research. First, even though spoken language is one of the most important means for expressing and conveying emotion-related information, there is scant research on the effects of the lexical meaning of spoken words on human emotions. Thus, it is important to study how the lexical content of spoken words affects emotional responses in humans. Second, because emotions have been recognized as an important part of human–computer interaction (HCI), it is essential to examine how people respond to the emotional expressions of computers. Third, because interfaces that utilize speech synthesis have become increasingly popular, it is important to understand which kinds of emotional reactions synthetically produced speech induces in people.

This thesis summarizes five publications that investigated how the lexical emotion-related content of synthesized speech affected people's emotion-related experiences and physiological responses in terms of facial-muscle and autonomic nervous system activity (i.e. pupil size and heart rate changes). In addition, the effects of emotional content on the perception of the quality of speech synthesis were studied.

First, the results showed that the emotional messages (i.e. sentences and words) produced by synthesized speech had significant effects on people's physiology and the ratings of their emotional experiences. Thus, passive listening to verbal emotional material induced changes both on the subjective and physiological levels. Second, the results provided evidence that the human-likeness of synthetic voices matters in respect to emotions.

Generally, more human-like synthesizers evoked stronger ratings for emotions than less human-like synthesizers. Further, comparisons between less and more human-like voices showed that only the more human-like synthesizers evoked significant emotion-related facial-muscle and pupil responses. Third, the results highlighted how the content of a message affected how people perceived and rated the speech synthesizers. When the content of the message was positive, the participants rated the voice as more pleasant and clear than when the content was negative or neutral.

In summary, the results presented in this thesis suggest that the synthesized lexical expressions of emotions can evoke emotions in people. This finding indicates the importance of language in human communication. Even though the spoken stimuli were generated by the monotonous voices of speech synthesizers and lacked interaction context, the stimuli activated the human emotional system. However, the features of the voice also matter when evoking emotions through computers. Finally, the results showed that the lexical content of the messages had such a strong effect on people that the impression of the voice quality was affected by the content of the spoken message. Previous research has suggested that interaction with computers is intrinsically social; consequently, people tend to use similar interaction rules both in HCI and in human–human interaction. Overall, this thesis finds that computers also evoke emotionality in their users. It seems that the emotional expressions produced by computers could evoke similar emotional responses in humans as the emotional expressions of another human could.

Acknowledgements

There are several persons to whom I would like to express my gratitude. First, I want to sincerely thank my supervisor, Professor Veikko Surakka, who has taught and encouraged me. I am grateful for his invaluable support and guidance throughout my work with this thesis. His enthusiasm for science is contagious.

This study was carried out at the Tampere Unit for Computer-Human Interaction. The research environment and facilities for preparing this thesis were excellent. Most of the funding for my thesis came from the Doctoral Program in User-Centered Information Technology (UCIT). This work was also financially supported by the Academy of Finland, the Jenny and Antti Wihuri Foundation, and the University of Tampere. I thank the reviewers Professor Timo Saari and Docent Martti Vainio for their time, effort, and valuable comments on this thesis.

It has been great to work at the Research Group for Emotions, Sociality, and Computing. I would like to thank all past and present members of the group; the pleasant and friendly atmosphere has made the work a great pleasure. More specifically, I wish to thank Outi Tuisku, and all those with whom I have shared office space: Jani Lylykangas, Jenni Anttonen, and Henna Heikkilä, for discussions, assistance, and advice. Special thanks also go to Toni Vanhala, who helped me with several data-processing issues. I would like to thank the helpful administrative staff. Thanks also to Tuula and Minna for the nice lunchtime company.

I am thankful to Erja and Heidi, and also to my other friends for the mental support and for sharing the moments of joy and frustration. My family has always been the pillar of my life. I want to thank my brother Tatu for being there. My father has always believed in me, encouraged, and supported me to realize my goals. My loving thanks to my husband, Pasi, and the sunshine of my life, our children, Tino and Sanja. Without you, every accomplishment would be empty. Words cannot describe how much you mean to me.

Last year was an extremely hard time in my life. I dedicate this thesis to the memory of my dear mother whom we lost to cancer in the fall when I was finalizing this thesis. I miss her every day. I am grateful for her unconditional love and support, and I am glad that she was able to see this process coming to its completion.

Tampere, 22th of May, 2013, *Mirja Ilves*

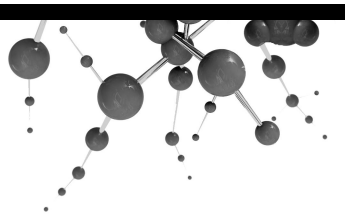
Contents

List of publications

This thesis consists of a summary and the following original publications, reproduced here by permission.

Author's Research Contributions

Each publication included in this thesis was coauthored, indicating that all of them originated from collaborative research between the authors. The research involved collaboration between the authors at every stage, but the author of this thesis was the main contributor to the design of the empirical work, the conduction of the empirical work, the analysis of the collected data, and writing the publications. The author of this thesis worked as the experimenter (i.e. performed the laboratory work), conducted data analysis, and wrote the first draft of each publication.



1 Introduction

Spoken language is one of the most important means of human communication. For example, speech is an effective way to convey information concerning one's intentions, ideas, and emotions to other people. Further, speech is the most important means by which (hearing) people build social connections and relationships with other people. From a historical perspective, the emergence of human language was one of the major transitions in human evolution (Maynard-Smith & Száthmáry, 1997). Along with the development of language, human communication was no longer only restricted to the present moment, as people could also speak about past and future events. Further, language enabled people to represent and communicate information that was complex and abstract (Fedurek & Slocombe, 2011). Thus, language made more diversified communication and interaction between people possible. Language is one of the capacities that distinguish humans from other species, and it has made human culture, with all its achievements, possible (Miller, 1981).

Nowadays, human communication is not restricted to other humans as people can also communicate with technology. People carry out many tasks with the help of different kinds of technological apparatus, and it is important to consider how the communication between humans and these technological devices could be improved. Human–computer interaction (HCI) is a research area that focuses on studying and designing the interactive elements between people and computers. A basic idea behind the field of HCI has been to make computer use more natural, effective, and enjoyable. One possible approach toward this goal is to design interfaces so that they mimic the ways in which humans communicate with each other. Due to its unique and central role in human communication and interaction, speech as an interaction means between

humans and computers has attracted researchers' interest. For several decades, researchers have developed and tested systems that can recognize spoken words, and they have developed and constructed systems that can produce speech. Interfaces that utilize speech have become increasingly common over the last few decades. For example, applications for visually impaired persons, applications in mobile phones, navigation systems, and telephone services all use speech output. One aim of the synthesized spoken messages is to communicate factual content, but spoken messages can also be used to direct and change the behavior of a computer user.

Emotions are one of the most central motivators of human behavior. There is evidence that emotions significantly affect various cognitive processes such as perception, memory, and thinking. Damasio (1994) has even shown that intact emotional processing is a necessity for rational decision-making. Not only are emotions and cognition interconnected, but emotions have an important role in social interaction and communication. Emotions that arise in human–human communication guide further interaction (Zajonc, 1980). In addition to other humans, machines and computer systems also evoke emotions in people. We all know that people can become attached to various objects and machines, and get angry with a computer if it does not work properly, for example. There is much scientific evidence that people react in much the same way to technology as they do toward other people, and use similar interaction rules when interacting with a computer as they do with other humans (Reeves & Nass, 1996). These findings relating to the importance of emotions in both communication and human rational behavior have also affected the recognition of the importance of emotions in HCI. The special field in HCI that aims at integrating emotions into HCI is called affective computing. Rosalind Picard published a pioneering book called *Affective Computing* in 1997, where she defined affective computing as "computing that relates to, arises from, or deliberately influences emotions." Basically, affective computing studies the issues relating to how computers can recognize, interpret, and represent emotions.

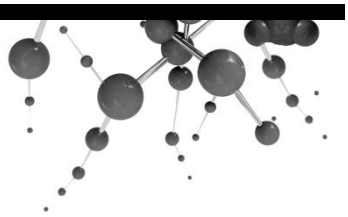
When I started my work, there was quite recent evidence that people reacted to voice technologies in a fairly similar way as they would to other people (e.g. Nass, Moon, & Green, 1997; Nass & Lee, 2001). Evidence was found that emotional information delivered through speech synthesis could have significant effects on human emotions and cognitive performance during computerized problem-solving tasks (Aula & Surakka, 2002; Partala & Surakka, 2004). The results of these studies awakened my interest in studying the potential of speech synthesis in conveying emotional information. Some studies had shown that computer systems could be programmed to express some emotions, for example, through the facial expressions of computer agents or through the nonverbal cues of

synthetic speech so that people recognized them (e.g. Bartneck, 2001; Murray & Arnott, 1995). However, it was not yet clear how people responded to the emotional expressions of computer systems. Thus, there was an apparent need to study what kinds of experiences computer-generated emotional information could evoke in people, and, particularly, if synthetically created emotional expressions could evoke emotions in people.

In spoken verbal expressions, emotions can be expressed in two ways. One way is through the lexical meaning of speech, and the other is through nonverbal information, such as loudness, the rate of the speech, the averaged fundamental frequency (F0), or range of the F0 of the voice. Previous studies have concentrated mainly on studying speech prosody, which is related to emotions. However, when people communicate their emotions through speech, the lexical meaning of the words is at least as important as speech prosody. Thus, it is important to also study how the lexical content of the spoken words affects the emotional responses of humans. Speech synthesizers offer a good means with which to study the effects of the purely verbal content of spoken words on the human emotion system. That is because they offer good controllability over timing and prosodic cues related to the nonverbal expressions of emotions. Controlling for the variation in nonverbal cues offers the possibility of being more able to study purely the effects of the lexical meaning of spoken words.

Synthesized speech can be created by using different techniques. The output of the different synthesizers varies from machine-like to very realistic and human sounding. Thus, speech synthesizers offer a good opportunity to also study the effects of the naturalness or human-likeness of voice on human emotions. Previous studies about the human-likeness of computer agents have been somewhat contradictory, and the studies have mainly only concentrated on studying the effects of computer agents' exterior features. Moreover, there is very little knowledge on how the human-likeness of a computer specifically affects people's emotion-related responses.

In sum, the main interest of this thesis is in studying if emotion-related responses are evoked in people by synthetically produced spoken messages with emotional content. The focus is on examining how the lexical emotional content of speech affects the emotions of a listener; that is, how the lexical content of speech activates the human emotional system. The other aim is to investigate how the human-likeness or naturalness of the voice affects responses. More specifically, does the emotional message delivered through a more human-like voice enhance the effects of the message?



2 What Is an Emotion?

Emotions are complex phenomena that have attracted the attention of researches for a long time. Over a hundred years have gone by since William James (1884) wrote his famous article "What Is an Emotion," yet there is still no common shared definition of *emotion*. According to Kleinginna and Kleinginna (1981), over 90 different definitions of emotion exist. However, it is generally agreed that emotions have a short duration, lasting from a few seconds to a few minutes at most, and that emotions usually have a specific object of focus. These features separate emotions from other closely related concepts and terms. It has been suggested (e.g. Gross, 2010) that affect can be considered as a more general concept, and that attitudes, moods, and emotions are different types of affective states (Figure 1). In contrast to emotions, moods can be considered as background feelings with a relatively long duration and lower intensity level, whereas one's emotional attitude is a general pattern of how one perceives the world.

The different theories of emotions approach emotions from different angles, emphasizing, for example, evolutionary, physiological, neurological, or cognitive aspects of emotions. From a historical perspective, it is possible to highlight some influential researchers who have emphasized these aspects differently, and who have all significantly influenced current research in the field (Plutchik, 1980). In the early history of the psychology of emotion, before James' (1884) seminal work, Darwin published a book named *The Expression of the Emotions in Man and Animals* (1872/1965). Darwin concentrated on studying emotionally expressive behavior, and he proposed that expressive behavior was the outcome of evolutionary development. According to him, there were some common features between the expressions of animals and humans. Thus,

he concluded that expressions had served some functions of communication, and had thereby affected the chances of survival.

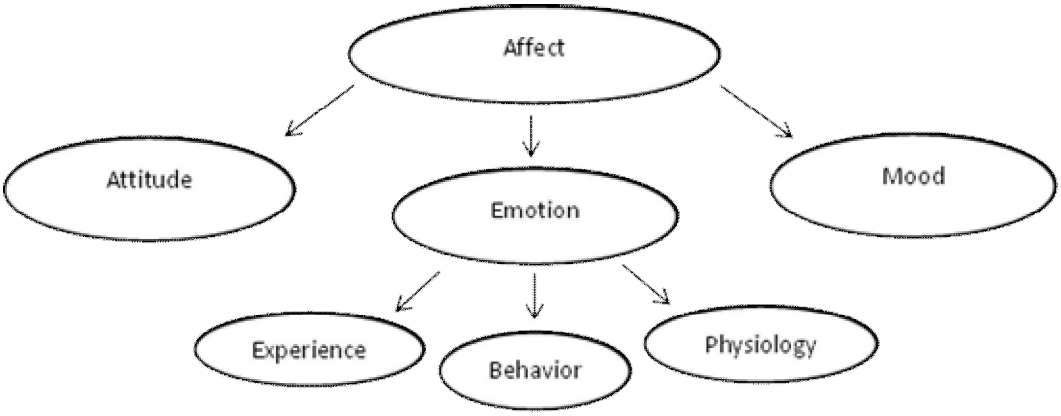


Figure 1. One possible way to organize key terms in affective science. Redrawn and slightly modified from Gross (2010).

While Darwin can be considered as a pioneer of the evolutionary view of emotions, the theory of William James has contributed to the psychophysiological tradition of emotions (Plutchik, 1980). James (1884) was mainly interested in the relationship between bodily changes and subjective feelings. James suggested that emotions originated from the bodily changes that followed from the perception of some fact or event. According to him, human bodies have evolved to respond automatically to those features of the environment that have survival-related significance to us. Our bodies respond first, and our perception and experience of these changes constitutes something that we call an *emotion*. For example, we do not cry because we feel sad, but we feel sad because we cry. For James, perceptions evolving from the autonomic nervous system (ANS) activity were central to emotions, and he suggested that different ANS responses result in different emotions.

According to another American professor, Walter Cannon (1929), one problematic aspect in the theory of James was that it assumed that each different emotion produced a unique set of bodily changes (Plutchik, 1980). Cannon frequently found evidence in his own studies that was not in accordance with James' theory. Among other things, he found that the lack of feedback from the ANS did not prevent the emotional expressions of animals. On the other hand, different negative emotions seemed to evoke very similar physiological reactions. Cannon was the first researcher to emphasize the importance of the brain in human emotions. He proposed a theory that the activity of the lower parts of the brain simultaneously caused an emotional experience and physical changes.

According to him, physiological correlates that differentiate between different emotional states emerge in the brain, resulting from the peripheral physiological responses.

One of the earliest studies that investigated the importance of cognitive interpretation on emotional state was the experiment by Schachter and Singer (1962). They suggested that, even though physiological arousal is connected to emotions, each particular emotion does not have its own pattern of response. Instead, they suggested that the cognitive appraisal of the situation may affect how people interpret their physiological arousal. Thus, an emotional reaction can result from the interpretation of the situation. In line with the suggestions from Schachter and Singer, current appraisal theories emphasize the importance of cognitive factors on emotions. These theories suggest that emotions result from the individual's evaluations and interpretations of their circumstances (Ellsworth & Scherer, 2003). In the appraisal process, a person evaluates physical, social, and mental changes in the environment mirroring their relevance to the current goal, and assesses the controllability of the stimulus. Frijda (1986; 1989) has suggested that the central element in emotion is a change in action readiness that is elicited if some event is appraised as emotionally relevant and somehow important to people. Action readiness means that the motivational state guides future action and physiological changes. Generally, theorists presume that appraisal is an automatic and unconscious process (Ellsworth & Scherer, 2003). However, Ochsner and Feldman Barrett (2001) have proposed that there are two forms of appraisal. On the one hand, there is automatic emotion processing, which enables the rapid and automatic detection of emotionally relevant information. On the other hand, we can deliberately monitor, activate, or process emotions, and thus, consciously regulate and reconstruct the meaning of an experience.

In sum, several different theoretical approaches to emotions exist. The different theories describe and try to explain the different aspects of emotions, and none of them is recognized as dominant. However, on a higher level of generality, it is possible to divide the different emotion theories into two groups, depending on whether emotions are discretely or dimensionally classified. The discrete view of emotion defines emotions as separate, distinguishable categories, whereas the dimensional view proposes that emotions can be defined through a certain set of dimensions. These two approaches are briefly introduced in the next section.

2.1 Discrete and Dimensional Theories of Emotion

Discrete emotion theories (e.g. Izard, 1977; Ekman, 1992; Johnson-Laird & Oatley, 1989) are based on the idea that a set of discrete primary emotions exists that is also termed as the *basic emotions*. According to discrete

theories, every primary emotion has its own neural and physiological background, and all other emotions can be defined through these primary emotions. However, no consensus exists among researchers on how many basic emotions there are or what those basic emotions are (Ortony & Turner, 1990). One of the most influential theories on discrete emotions is Paul Ekman's basic theory of emotions, which proposes that basic emotions also have distinctive, universal facial expressions (Ekman, 1994). Ekman (1992) has suggested nine characteristics, which are common to all basic emotions, and, on the other hand, that differentiate emotions from one another, and, for example, from moods. These characteristics are:

1. Each emotion has a distinctive, universal facial expression.
2. It is possible to find comparable expressions in animals.
3. Emotions have a distinctive pattern of ANS activity.
4. It is possible to find universal antecedent events for emotions.
5. There is coherence in expressions and autonomic changes during emotions.
6. Emotions can begin quickly before conscious awareness.
7. The duration of emotions is brief, usually lasting for seconds.
8. The appraisal mechanism before an emotional response is automatic, and it occurs quickly and without awareness.
9. The occurrence of emotions is unbidden.

Later on, Ekman (1999) added two more characteristics:

10. Emotions regulate thoughts, memories, and expectations.
11. Every emotion has a distinctive subjective experience.

Ekman (1992) proposed six basic emotions (joy, fear, anger, disgust, sadness, surprise), which share the above-mentioned characteristics. Every one of these basic emotions can be seen as an emotion family rather than as a single affective state. For example, joy describes a family of positive emotions, which includes different types of positive emotions, such as amusement, excitement, and satisfaction. In addition to basic emotions, there are also other more complex emotions, which can be derived from the basic emotions, or which are some sort of combination of the basic emotions (Ekman, 1992; Johnson-Laird & Oatley, 1989).

Dimensional theories of emotions define emotions through a set of dimensions. The basis for the development of the dimensional emotion theory can be found in the work of Wilhelm Wundt. Wundt (1896) proposed over a hundred years ago that three dimensions—pleasantness, excitation, and relaxation—could describe human feelings. Later on, Schlosberg (1954) continued the work and suggested that it was possible to describe emotion expressions through a set of dimensions instead of treating emotions as discrete states. His three dimensions were the level of

activation, pleasantness–unpleasantness, and attention–rejection. Some decades later, Russell's (1980) circumplex model of affect proposed that all emotions could be organized through two bipolar dimensions: pleasure–displeasure and arousal–sleep. This implies that every affective state consists of both pleasantness and activation components (Yik, Russell, & Feldman Barrett, 1999). The pleasure–displeasure (or valence) dimension refers to the hedonic tone of experience, and the activation (or arousal) dimension refers to the sense of mobilization or energy (Feldman Barrett & Russell, 1999). Several further studies have concluded that both of these bipolar, independent dimensions are necessary and sufficient to describe the *core affect*, although other dimensions have also been suggested (Feldman Barrett & Russell, 1999; Russell & Feldman Barrett, 1999). Russell and Feldman Barrett (1999) have suggested that the core affect is the most elementary affective feeling that is consciously accessible. According to them, the core affect is at the heart of any emotional episode: it is always present, but it varies in intensity. Whereas specific emotions arise from the interpretation of the neurophysiological changes in the valence and arousal systems (Posner, Russell, & Peterson, 2005).

Bradley and Lang (e.g. Lang, Bradley, & Cuthbert, 1992; Bradley & Lang, 2000a) have further developed the dimensional theory of emotions. They suggest that human emotions are organized along three dimensions: valence, arousal, and dominance. Among these three dimensions, valence and arousal are the most frequently studied dimensions. The valence dimension represents the pleasantness of an emotional experience, ranging from an unpleasant to a pleasant experience. The arousal dimension refers to the intensity of activation, ranging from an unaroused state to high arousal. Bradley and Lang (e.g. Lang et al., 1992) suggest that the valence and arousal dimensions represent underlying appetitive and aversive motivational systems of the brain that guide people in approaching or withdrawing from different stimuli.

Both the discrete and dimensional views have quite strong empirical support, and it is possible to see these two theoretical views as complementary rather than contradictory (Christie & Friedman, 2004). For example, it is possible to locate discrete emotions in a specific place within a dimensional affective space. However, in practice, researchers must usually choose either the discrete or the dimensional approach as a basis for their research framework. The dimensional approach was chosen as a basis for the experimental research of the present thesis, because it seemed more suitable at the early stage of studying the relationship between the synthesized lexical expression of emotions and human emotions than the discrete approach. The use of bipolar dimensions enables a freer evaluation of emotional experiences than when reporting emotions using a pre-defined list of specific emotions. Bipolar dimensions also provide information concerning the degree of pleasantness and arousal, and not

only information about a specific emotional experience. For example, fear does not always consist of an identical amount of displeasure and arousal, as the relative amounts vary from situation to situation. In addition, there are many standardized sets of affective stimuli that have been studied in the context of the dimensional approach. One of the largest and most widely used word sets is the affective norms for English words (ANEW), which provides the normative ratings for valence, arousal, and dominance for over 1000 words (Bradley & Lang, 1999). Thus, it provides for a good possibility of selecting word stimuli so that they vary systematically on the emotional dimensions. Finally, there is evidence that emotional dimensions correlate well with physiological changes (Mauss & Robinson, 2009).

2.2 Measuring Emotions

Although there is no unanimous definition of an emotion, researchers agree quite widely that during an emotional event, three different components of human behavior are activated. These include changes in physiology, experience, and behavior (e.g. Izard, 1977; Frijda, 1986). First, emotions produce changes in the neural system. Emotions also cause changes in subjective experiences; that is, people feel happiness or anger, for example. Finally, emotions affect human expressive and motor behavior, for example, so that an unpleasant situation may lead to withdrawal, while a pleasant situation may lead to approach behavior. As emotions cause changes on multiple levels, studies relating to emotions often apply measurements to many levels at the same time. Using multiple methods makes it possible to gain information about the different aspects of the emotional reaction. The different measurement methods can be divided, for example, into two categories, depending on whether they require participants to self-evaluate and report their emotions or not.

Methods concentrating on some kind of self-evaluation about one's own emotions consist of different types of self-report evaluations of emotional experiences. In addition to measuring and analyzing participants' ratings of their own possible emotions, the use of self-report ratings for emotions are a necessity because they are also required to confirm that the used stimuli affected people's emotional experiences as expected beforehand. It is noteworthy that the use of these methods is the only method used to gain information about the subjective experiential component of emotion. The use of self-evaluation for experienced emotions seems a quite simple and straightforward method to collect information about emotions. Although at first sight these methods seem to be simple and easy to apply, quite a few problems exist in relation to them. For example, Robinson and Clore (2002a, 2002b) have suggested that, at best, self-reports give important information about current experiences, but it is not possible to store or retrieve the emotional experience. By this, they mean that a delay

between the experience and the reporting of the experience causes a loss of information, and that reactivating a past situation does not result in the same experience as the original situation, but it is a new emotion created in the present moment. Thus, to get valid information about emotional experiences, it is important to collect the emotional ratings of current experiences instead of past emotions that are not currently being experienced (Robinson & Clore, 2002b). In many cases, the use of different rating scales also interferes with the experimental settings, so that during or in-between different stimulations one has to try to cognitively analyze what has happened on the emotional level. This, of course, can interfere in many ways with evoked emotional processes as well.

A wide variety of tools has been used to measure emotional experiences. They range from checklists where a respondent indicates whether she/he experienced some specific emotion, to the use of unipolar or bipolar dimensional scales (Larsen & Prizmic-Larsen, 2006). Today, it is widely agreed that two earlier described dimensions—valence and arousal—are the most primary and most widely used emotional dimensions (Bradley & Lang, 1994; Mauss & Robinson, 2009). The valence dimension ranges from an unpleasant to a pleasant emotion, and the arousal dimension ranges from a calming to an aroused feeling. Bradley and Lang (1994) have developed a pictorial rating technique called the self-assessment manikin (SAM). For example, the valence dimension in SAM varies from a smiling figure to a frowning figure. There are five pictures for each scale, but a participant can also rate his/her experiences by making a mark between any two figures, which results in a nine-point scale. In all the experiments for this thesis, the numerical versions of the SAM's valence and arousal scales were used when the participants rated their subjective emotional experiences. The scales were nine-point bipolar scales ranging from one to nine. The valence scale ranged from a negative experience to a positive experience, and the arousal scale from a calm experience to an aroused experience. The center of both scales represented a neutral point (e.g. neither an unpleasant nor a pleasant experience). 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 the dimensional theory of emotions suggests that, importantly, the approach-avoidance tendency has a central role in emotional processing, it has not been measured frequently. Recently, ratings for this tendency have been initiated by asking about the ratings for the approach-withdrawal tendency using a bipolar rating scale varying from avoidable to approachable (e.g. Anttonen & Surakka, 2005).

The use of self-reports has been criticized because they are dependent on the participants' ability and willingness to report their emotions (Larsen & Prizmic-Larsen, 2006). For example, people may try to answer in a socially desirable way, or people are not always aware of the changes in their

emotional state. The processing of emotional information is fast, and even subliminally presented emotional stimulations can evoke physiological responses and changes in subjective ratings (e.g. Zajonc, 1980; Whalen, Rauch, Etcoff, McInerney, Lee, & Jenike, 1998). In contrast to subjective experiences, which cannot be tracked in a totally continuous manner, physiological changes can be measured continuously on a moment-by-moment basis. In addition, physiological measurements do not have to interrupt participants through intervening tasks, thus giving potential access to continuous, emotion-related physiological changes.

Changes in physiology caused by emotions include both central and peripheral nervous-system changes. The central nervous system consists of the brain and the spinal cord. The peripheral nervous system consists of two subsystems, which are the somatic nervous system and the ANS. The primary function of the somatic nervous system is the voluntary control of the skeletal muscles. The ANS primarily regulates involuntary activity, such as the heart rate, pupil size, and skin conductivity. The ANS is typically further divided into two main subsystems: the sympathetic nervous system and the parasympathetic nervous system. The sympathetic nervous system activates bodily changes, preparing people for the fight or flight response; that is, to defend or to escape. The parasympathetic nervous system relaxes bodily activation, returning the activation to a normal or baseline level. Most organs receive impulses from both subsystems, and thus, the current balance between these systems defines if the effects of the ANS activity are inhibitory or excitatory. Of course, there are also issues with physiological measurements, such as with the presence of artifacts or with problems concerning the signal-to-noise ratio. In any case, subjective ratings of emotional experiences have an important role when interpreting the results of the measurements of physiological reactions. Next, I will examine studies that provide a background on how measurements that reflect physiological processes can be related to human emotions.

There is some evidence that different discrete emotions have a distinct ANS activity pattern (e.g. Ekman, Levenson, & Friesen, 1983; Levenson, Ekman, & Friesen, 1990; Levenson, 1992; Christie & Friedman, 2004; Rainville, Bechara, Naqvi, & Damasio, 2006). In one of the earliest influential studies, Ekman et al. (1983) instructed their participants to produce the facial expressions of emotions or to recall past emotional experiences, and then measured their ANS activity. They found some differences in autonomic activity (i.e. heart rate, finger temperature, skin resistance) between positive and negative emotions, but also among some negative emotions. More recently, Rainville et al. (2006) found that the experience of fear, anger, sadness, and happiness were associated with distinctive patterns of cardiorespiratory activity.

Even though there is some evidence for the different autonomic responses between some emotions, the results regarding autonomic specificity are far from definitive (Cacioppo, Berntson, Klein, & Poehlmann, 1997; Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000). Mauss and Robinson (2009) concluded—based on their recent review—that the dimensions of valence and arousal seem to capture the variance in physiological responding better than discrete states do. When studying physiological responding in the framework of dimensional emotion theory, interest is focused on the relationship between the physiological changes and the valence and/or arousal dimensions.

Heart rate is one of the most commonly used physiological signals reflecting the ANS activity that is associated with experienced valence. Basically, there are two types of findings. First, several previous studies have reliably shown that heart rate decelerates during the viewing of emotional stimuli, so that the deceleration is the largest and most prolonged response to negative stimuli. This pattern of response has been obtained for emotional pictures (Bradley, Cuthbert, & Lang, 1996; Palomba, Angrilli, & Mini, 1997), sounds (Bradley & Lang, 2000b), and dynamic facial expressions (Anttonen, Surakka, & Koivuluoma, 2009). There is also some evidence that verbal material evokes similar kinds of heart-rate responses. Buchanan, Etzel, Adolphs, and Tranel (2006) found that heart-rate deceleration was greater for visually presented, unpleasant words than for taboo, school-related, or neutral words. Second, there is evidence of accelerating heart-rate responses. For example, when people are instructed to remember or imagine emotional material, or to produce voluntary emotional facial actions, their heart-rate responses generally accelerate (e.g. Ekman et al., 1983; Lang, Bradley, & Cuthbert, 1990; Waldstein, Kop, Schmidt, Haufler, Krantz, & Fox, 2000). In sum, the heart rate tends to decelerate during stimulus intake when the attention is directed to external stimuli, and, on the other hand, to accelerate when the task requires more inward processing.

When people perceive emotional stimuli, their heart-rate changes typically follow a triphasic form consisting of an initial deceleration followed by acceleration, and then a late deceleration (Lang Bradley, & Cuthbert, 1997; Bradley & Lang, 2000a). Overall, the heart rate decelerates more in response to unpleasant than to pleasant or neutral stimuli. The defense cascade model (Lang et al., 1997; Bradley & Lang, 2000a) has explained this deceleration response. According to the model, any new stimulus in a viewing context causes a heart-rate deceleration indexing an orienting response. The stringer cardiac response to unpleasant stimuli compared to pleasant or neutral stimuli suggests the pronounced allocation of attentional resources toward stimuli that are interpreted as somehow threatening. Heart-rate deceleration turns into heart-rate acceleration only at the highest level of action, just before action, but usually this does not

happen when processing standard unpleasant stimuli (Bradley & Lang, 2000a).

Changes in the level of skin conductance and pupil size are other examples of the physiological signals that result from the activity of the ANS. There is evidence that changes in these signals relate especially to emotional arousal. Skin conductance reflects the changes in sweat gland activity and it is a sensitive measurement associated with the novelty of the stimulus and emotional arousal (Dawson, Schell, & Filion, 2000; Bradley, 2009). Several studies have found that skin conductivity correlates with arousal, so that skin-conductance responses are greater for arousing negative and positive stimuli than for neutral stimuli.

In comparison to skin conductance, pupil size is a more rarely studied signal. The pupil is the round black opening in the center of the eye, which regulates the amount of the light entering the eye. There is also evidence that changes in pupil size are associated with mental activity. Previous studies have shown that pupil dilation correlates with cognitive load. The more mental processing effort is needed, the more the pupil dilates (e.g. Schluroff, 1982; Verney, Granholm, & Dionisio, 2001; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004). In addition, several studies have shown that the pupil dilates in response to emotional information. Even though the findings of pupil-size variation related to affective processing have been somewhat controversial, most studies have consistently suggested that pupils dilate during emotionally arousing conditions, whether the condition is pleasant or unpleasant. For example, there is evidence that the pupil dilates more during viewing or listening to a pleasant or unpleasant stimulus as compared to neutral stimulation (Janisse, 1974; Partala & Surakka, 2003; Bradley, Miccoli, Escrig, & Lang, 2008)

In addition to autonomic activity, the activation of facial muscles can provide information about emotions. There is much evidence that facial-muscle activations are related to experienced valence. The Facial Action Coding System (FACS) is a widely used technique for objectively describing facial activations (Ekman & Friesen, 1978). It separates out 46 action units from the human face, which are related to the activity of facial muscles or a muscle group. Sometimes emotions are so weak that they do not evoke facial actions that are visually perceptive. Therefore, facial electromyography (EMG) offers a methodology with which to study emotional processes and emotions that are too weak to be visible to facial action coding (Cacioppo et al., 1997). The two most studied facial muscles are the corrugator supercilii and zygomaticus major. The corrugator supercilii is a facial muscle that is responsible for the lowering and contraction of the eyebrows. The zygomaticus major is a facial muscle that is responsible for smiling. It has been found that the activation of the

zygomaticus major increases during positive stimulation (e.g. Lang, Greenwald, Bradley, & Hamm, 1993). Conversely, several studies have provided evidence that the activity of the corrugator supercilii correlates linearly with experienced valence, so that the activity of the corrugator supercilii increases during negative emotions, and relaxes during positively valenced emotions (Lang et al., 1993; Bradley & Lang, 2000b; Larsen, Norris, & Cacioppo, 2003). In addition to emotional valence, corrugator activity may also be influenced, for example, by cognitive factors such as concentration (Cacioppo, Petty, & Morris, 1985).

Finally, the dimensional theory of emotions suggests that the dimensions of valence and arousal are connected to the motivational system of approach or withdraw. One line of evidence comes from neurophysiological studies. Different brain-imaging methods such as electroencephalography (EEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) have shown that several separate brain regions are involved in emotional processing (see reviews by Phan, Wager, Taylor, & Liberzon, 2002; 2004). Some areas may be especially important for some emotions, as for example, the amygdala has an important role in fear-related processing, whereas some other areas, for example, certain regions in the cortex, are more involved in certain types of emotional tasks regardless of what the specific emotional state is. On the other hand, EEG studies have provided evidence that there seems to be hemispheric asymmetry in the prefrontal activation relating to the motivational tendency to approach or to avoid. Among the first to demonstrate this was Davidson (1992), who suggested that approach-related (i.e. positive) emotions evoke more activation in the left frontal regions, and withdrawal-related (i.e. negative) emotions evoke more activation in the right frontal regions. More recently, Coan, Allen, and Harmon-Jones (2001) monitored the brain activity of participants while they produced emotional facial expressions. Consistent with the findings of Davidson (1992), they found that there was relatively less left frontal activation during the withdrawal condition than during the approach condition. Overall, the evidence supports the view that cortical areas are related to motivational tendencies rather than to the valence of affective experience. Approach–avoidance motivation is often associated with valence, but not always. For example, anger, which is a negatively valenced emotion, relates to an approach-oriented motivational system (Carver & Harmon-Jones, 2009).

In sum, emotions consist of different components that can and have been measured using different methods such as asking for ratings of emotions or measuring the changes in physiology. However, it is not possible to measure many of these changes simultaneously; thus, researchers have to select a small set from the different measures. In this study, the focus was mainly on subjective ratings. Some studies also measured heart-rate

changes, pupil size, and facial electromyography to elucidate what kinds of physiological responses synthesized spoken messages evoke in listeners.

2.3 The Mediator Role of Emotions in Cognitive Processes and Well-Being

It was long thought that cognition and emotion were two separate mental processes. Damasio (1994) was among the first to show that cognitive processes were at least partly dependent on emotions. He found that patients with brain damage in the areas that are involved in emotion processing had difficulties in rational decision-making. Several later studies have supported the view that emotions also affect, in addition to decision-making, other cognitive processes, such as creative thinking, problem solving, attention, memory, and motivation.

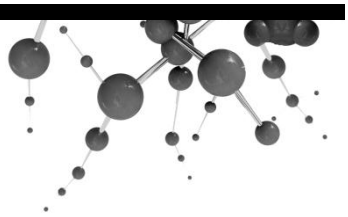
Overall, there is general agreement that even a mild positive affect enhances cognitive performance during certain types of tasks (Ashby, Isen, & Turken, 1999). For example, the studies of Isen and colleagues have shown that the positive affect that was induced by giving a small bag of candy or by showing a few minutes of a comedy film can improve performance for tasks that require creative problem solving (Isen, Daubman, & Nowicki, 1987; Estrada, Isen, & Young, 1994). Positive emotions can also promote creativity, increasing the originality of ideas (Grawitch, Munz, Elliott, & Mathis, 2003). On the other hand, there is some evidence that positive and negative emotions elicit different types of problem-solving strategies. Spering, Wagener, and Funke (2005) studied how positive and negative emotions affect performance in complex problem solving. They found that positive or negative emotion did not affect overall problem-solving performance, but instead, positive and negative emotions elicited different problem-solving strategies. Participants with negative emotions had more detailed and systematic ways of seeking and using information than participants with positive emotions.

There is also evidence that emotions affect perception and memory. Usually, people perceive and remember information that has emotional content better than non-emotional information (e.g. Yiend, 2010; Doerksen & Shimamura, 2001; Kensinger & Corkin, 2003). Further, Fredrickson and Branigan (2005) found that positive emotions can broaden the scope of attention in relation to the neutral state. There is also evidence that positive emotions can facilitate the motivational state of people (Erez & Isen, 2002).

It has also been suggested that emotions can have effects on individuals' health and well-being. There is evidence that negative emotions (e.g. anger) play a role in the etiology of heart disease and some cancers (see e.g.

Fredricsson, 2000). Especially, individuals with poor anger-management skills have an elevated risk of cardiovascular diseases (Haukkala, Konttinen, Laatikainen, Kawachi, & Uutela, 2010). There is also evidence that higher frequency heart-rate responses last longer after negative emotions than after positive emotions (Brosschot & Thayer, 2003). These kinds of findings support the theory that the health risks of negative emotions are mediated, at least partly, by prolonged physiological activation (Brosschot & Thayer, 2003). In the long term, negative emotions may result in weakened immune-system functioning (Kiecolt-Glaser, McGuire, Robles, & Glaser, 2002; Rosenkranz, Jackson, Dalton, Dolski, Ryff, Singer, Muller, Kalin, & Davidson, 2003). On the other hand, positive emotions are associated with better physical health (Pressman & Cohen, 2005). In addition, positive experiences after negative ones can alleviate the effects of negative emotions. Fredrickson and Levenson (1998) found that positive emotions help an individual to recover from heart reactions that were caused by negative emotions.

In sum, several studies have shown that emotions have a central and important role in human cognitive performance, and in the long term, emotions can affect the health of an individual. Thus, it is conceivable that the role of emotions is important in HCI too. Working with computers is often stressful and cognitively demanding. By taking emotions into account, for example, by giving supportive feedback to a computer user, it might be possible to support the cognitive work of a computer user and, on the other hand, to reduce users' stress in the context of HCIs.



3 Speech and Emotions

As indicated earlier, language is probably the most important means when people communicate and build relationships, and it is a primary instrument for influencing others (Ng & Bradac, 1993). Specifically, language is an effective way to convey cognitive and emotional information to other people. Language communication occurs in spoken and written forms, and thus, it includes many kinds of lingual activities such as speaking, listening, reading, and writing. As compared to written language, spoken language is a more primary and more fundamental means of communication (McGregor, 2009). Every human society has a spoken language, even those without the tradition of writing, and most people hear and speak more than they write or read. In addition, historically, speech develops before written language. In every culture, speech has evolved before the written language, and in terms of individual development, speaking is learned much earlier and more easily than writing. Due to its importance in human communication, speech affords a great potential to develop human interaction and communication with computers.

In HCI, the use of speech can also have some advantages as compared to textual or other visual messages when the aim is to provide, for example, information, feedback, or encouragement to a computer user. Many current interfaces are overloaded with visual information. Thus, if we want to communicate a message from a computer to a user, visual information may be lost in the plenitude of other visual information. In addition, multiple resource theory suggests that in addition to the difficulty of tasks, the qualitative demands of the tasks affect dual-task performance (Wickens, 2002). For example, it is typically easier to divide attention between the visual and auditory channels than between two visual or two auditory channels. Further, Alais, Morrone, and Burr (2006)

have suggested that each sensory modality is, more or less, under a separate attentional control. Thus, when a user is carrying out some visual tasks, speech messages probably disturb visual processing less than additional visual messages.

Emotional information can be delivered through speech by using the verbal content of speech or prosody. Speech prosody refers to the nonverbal aspects of speech, such as loudness, the speech rate, F0, or the range of F0 of the voice (e.g. Scherer, 1986). To study purely the role of prosody or, alternatively, the role of verbal content on human emotions, it is important to keep the other role as neutral as possible. Studies concerning speech and emotions have mainly concentrated on studying speech prosody, and there is little evidence about the role of the verbal meaning of spoken words in human emotions. However, the results from the studies relating to emotions and visually presented linguistic stimuli can provide some background for studying responses to spoken emotional words, because emotionally charged spoken stimuli, uttered in a monotone or a neutral tone of voice, are somehow comparable to written text. That is, only the lexical content of the stimuli offer knowledge about emotion.

When we want to study purely how the verbal content of spoken words affects human emotions, speech synthesizers offer good opportunities, as they afford good controllability over the timing and prosodic cues (e.g. loudness or pitch) that can convey information about emotions. Neutralizing the variation in nonverbal cues gives us more of a chance to study purely the effects of the lexical meaning of spoken words.

Thus, this chapter initially presents an overview of the studies related to speech and emotions. As these studies are mainly concerned with studying speech prosody, studies relating to visually presented words can provide some background references for studying the lexical meaning of the spoken words on human emotions. Therefore, research related to written emotional information is introduced. At the end of this chapter, speech synthesizers are presented.

3.1 Prosody Cues versus Lexical Content

As mentioned, many studies have concentrated on examining which kinds of prosodic changes are related to emotions. One approach to studying emotion-related prosody is to analyze emotional speech samples. In these studies, actors or naïve subjects are asked to produce vocal expressions of emotions, or alternatively, material is recorded during naturally occurring emotional states. The findings across studies are quite consistent. For example, research has shown that values relating to F0 differ from one emotion to another (Banse & Scherer, 1996; Scherer, Banse, Wallbot, &

Goldbeck, 1991). On the other hand, knowledge regarding emotion-related speech prosody is still restricted and unsatisfactory (Bachorowski, 1999; Scherer, 1989). For example, there are some problems relating to speech samples. The simulated expressions of emotions that are produced by actors may not be natural enough, and on the other hand, speech samples from spontaneous speech may not be emotional enough (Scherer, 1986). The analyses of spontaneous speech samples have revealed that even though speech is rarely emotionally neutral, the emotions expressed in speech are rarely strong, pure emotions (Cowie & Cornelius, 2003).

On the other hand, decoding studies in which the participants' task is to infer the emotional state of a speaker from vocal expressions have revealed that people recognize emotions fairly well from the prosodic cues (e.g. Banse & Scherer, 1996). Some emotions are perceived more inaccurately than others are, but overall, the recognition accuracy is far better than chance (Scherer, 1986; Scherer et al., 1991). In addition, there is evidence that emotion-related psychophysiological responses can be evoked through the vocal expressions of emotions. The study by Hietanen, Surakka, and Linnankoski (1998) showed that the vocal expressions of anger increased the activity of the corrugator supercilii (i.e. active when frowning) more than the expressions of contentment did. In contrast, the activation of the orbicularis oculi (i.e. active during authentic smiling) was greater following expressions of contentment than following expressions of anger. Based on the findings, they suggested that anger and contentment may be contagious when *hearing* vocal affect expressions, and not only when seeing the facial expressions relating to the emotions.

As already noted, much less knowledge exists concerning the role of the meaning of spoken words in human emotional processing. Wexler, Warrenburg, Schwarz, and Janer (1992) studied the effects of negative and positive words on EEGs and facial EMGs in a dichotic listening paradigm. Both EEGs and EMGs provided evidence that even the stimuli that were processed without conscious awareness evoked emotion-specific responses. Bertels, Kolinsky, and Morais (2009) examined how an emotionally congruent or a neutral tone of voice affects the emotional ratings of words. Although the emotionally congruent tone of voice had a strengthening influence on some ratings, the words spoken by a neutral tone of voice also elicited emotion-related ratings of emotions, and, for example, the valence ratings of the positive words were not affected by the tone of voice.

There are also studies that have investigated the joint effects of verbal content and prosody. There are some contradictory findings as to whether an emotional tone of voice is connected to the perception and processing of spoken words. On the one hand, there is evidence that congruence or incongruence between emotional tone and word meaning does not affect

the processing of spoken words in lexical processing tasks (e.g. Wurm & Vakoch, 1996). On the other hand, there is evidence that words are named faster when the tone of voice is congruent with the semantic content of a word (Nygaard & Queen, 2008). Further, Schirmer, Zysset, Kotz, and von Cramon (2004) have studied the effects of gender on the processing of spoken words. In their research, men and women listened to positive and negative words spoken with congruent or incongruent prosody while their brain activity was measured by fMRI. They substantiated how emotional prosody affected the semantic processing of women more than the semantic processing of men. There is also some proof that there may be cultural differences involved regarding whether verbal content or vocal tone is more important when people process spoken emotional words. Ishii, Reyes, and Kitayama (2003) found that Western people (i.e. Americans) had greater difficulty ignoring the verbal content than ignoring the vocal tone of spoken emotional words, showing a bias for the verbal content of spoken emotional words in a Stroop inference task. Conversely, Asians (i.e. Japanese) had greater difficulty in ignoring the vocal tone than ignoring the verbal content, showing a bias for the vocal tone of spoken emotional words. Thus, they suggested that Western people have an attentional bias for verbal content, whereas Asians have a bias for the vocal tone of spoken words. In sum, the previous studies have shown that certain changes in speech prosody can convey information about emotions. However, research has also shown that the lexical content of spoken words can potentially have a stronger role in conveying emotional information.

3.2 Written Language and Emotions

Several studies utilizing event-related potential (ERP) measurements have shown that people identify the emotional content of written words at an early stage of lexical processing. Thus, it seems that the allocation of attentional resources to emotional words is automatic. For example, the studies have repeatedly found that visually presented emotional words are processed faster than neutral words (e.g. Herbert, Junghofer, & Kissler, 2008; Kissler, Herbert, Peyk, & Junghofer, 2007; Kissler, Herbert, Winkler, & Junghofer, 2009; Schacht & Sommer, 2009a; Scott, O'Donnell, Leuthold, & Sereno, 2009). Existing empirical evidence has also consistently suggested that emotional language can shape how a person perceives other emotional stimuli. De Houwer, Baeyens, and Eelen (1994) examined how subliminally presented positive and negative words affect the ratings of neutral words. They found that neutral words, which were followed by a subliminally presented positive word, evoked more positive feelings than neutral words that were paired with negative words. Halberstadt and Niedenthal (2001) found that faces expressing blends of happiness and anger were later encoded as angrier if the faces were originally presented with the word *angry* as compared to the same faces paired with

the word *happy*. These results suggest that by using linguistic information for emotions it is possible to shape how a person perceives other stimuli. Several quite recent studies have also shown that visually presented emotional words can evoke emotions in people.

This evidence comes mainly from research investigating ANS activity related responses. The studies have shown that ANS responses to written emotional words seem to be quite consistent with the reactions to affective pictures and sounds, yet it has been suggested that the reading of emotional words elicits a lesser degree of physiological arousal than when viewing affective pictures (e.g. Hinojosa, Carretié, Valcárcel, Méndez-Bértolo, & Pozo, 2009). Previous studies have found greater skin-conductance responses to written phobia-related, threatening, and aversive words as compared to neutral words, even when the words were not consciously identified (Silvert, Delplanque, Bouwalerh, Verpoort, & Sequeira, 2004; Van Den Hout, De Jong, & Kindt, 2000). Buchanan et al. (2006) also noted that visually presented emotion words elicit a similar pattern of autonomic response as previously found for emotional pictures and sounds. They showed that taboo words caused significantly larger skin-conductance responses than neutral words, and that heart-rate deceleration was the greatest and most prolonged in response to unpleasant words.

In addition to ANS activation, there is evidence that written emotional words elicit similar facial-muscle reactions as other emotional stimuli. More specifically, previous studies have found that written emotionally negative words evoke a stronger activation of the corrugator supercilii (frowning) facial muscle than positive words do (Bayer, Sommer, & Schacht, 2010; Foroni & Semin, 2009; Larsen et al., 2003). It has also been found that the startle reflex to unpleasant words is stronger than to neutral and positive words during shallow word processing (Herbert, Kissler, Junghöfer, Peyk, & Rockstroh, 2006; Herbert & Kissler, 2010).

Altogether, it seems that people's responses to visually presented emotional words are quite similar to their responses to other types of emotional information. However, the results of the responses to visually presented words are not directly applicable to spoken words. Thus, this thesis aims to investigate the role of the lexical content of spoken words in human emotional processing. Speech synthesis provides an excellent opportunity for that analysis.

3.3 Speech Synthesis

Speech synthesis means the artificial production of speech. Synthesized speech can be created through various techniques. Articulatory synthesis is a speech-synthesis method that attempts to generate speech by

mimicking the human speech-production system. The quality of articulatory synthesis is restricted, because there is not enough detailed and deep knowledge about the complex human articulatory system (Klatt, 1987). Two current primary techniques to create synthesized speech are formant synthesis and concatenative synthesis.

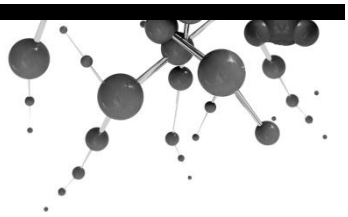
Formant synthesis generates speech by modeling the acoustic properties of human speech. The output sounds quite unnatural and robot-like, but the technique has some advantages over other synthesis methods. The computational requirements of formant synthesis are relatively small, and thus, it needs only a little amount of memory. Consequently, it is often suitable for systems where computational power and memory are restricted, such as in mobile applications (Iso-Sipilä, Moberg, & Viikki, 2006). In addition, it allows the detailed and rich control of speech signals, which makes formant synthesis interesting for modeling emotional expressivity in speech (Pierre-Yves, 2003; Schröder, 2009).

However, nowadays, the most widely used synthesis technique is concatenative synthesis, because its output is the most natural. In concatenative synthesis, speech is created by stringing prerecorded human speech sounds that are stored in a database. Concatenative synthesizers differ in the size of the used speech segments. The longer the speech units, the more natural, or human-like the voice sounds. However, a speech synthesizer that utilizes long units requires a large speech database, and thus, also more computational resources. Concatenated units can be small, such as in the microphonemic method, in which speech output is concatenated from approximately ten-millisecond-long samples of natural speech, or from much longer samples such as whole words. However, two main subtypes of concatenative synthesis are diphone synthesis and unit-selection synthesis. The database for diphone synthesis contains all the diphones (i.e. transitions from the middle of a phone to the middle of the next one) occurring in a language. For example, in English, there are about 1200 diphones (Chappell & Hansen, 2002). Unit selection is a technique utilizing a large database of recorded speech (i.e. several hours of speech), where the most appropriate units are selected. Recently, a fourth type of speech synthesis, a hidden Markov model (HMM) based speech-synthesis system has gained popularity. In HMM-based speech-synthesis systems, HMMs are trained on a speech database, and speech waveforms are generated from the HMMs themselves.

In the last few decades, there have been attempts to create emotionally expressive speech synthesis. Many existing systems have used formant synthesizers because they provide a high degree of control over acoustic parameters. There is evidence that by systematically manipulating the speech rate, pitch contour, and voice quality of synthetic speech, it is possible to generate emotion-like effects, which are recognizable to

listeners (Cahn, 1990; Murray & Arnott, 1995). By using concatenative synthesizers, it is possible to produce more natural-sounding speech, but the controlling of speech signals is more difficult with concatenative synthesizers than with formant synthesis. For example, in diphone-synthesis systems, usually only the F0 and duration can be controlled (Schröder, 2009). In systems that are based on unit-selection synthesis, one possible way to generate emotional speech is to use several different databases. Iida, Campbell, Higuchi, and Yasumura (2003) recorded a database that was spoken with angry, joyful, and sad tones of voice. To synthesize a given emotion, the units from the corresponding database were used. The results of the perceptual test showed that participants recognized emotions well (50–80%) from synthesized speech. Despite all the progress, emotionally expressive synthesis still has a long way to go (Schröder, 2009). For example, more detailed knowledge of the acoustic parameters relating to emotions is needed, and, on the other hand, better control over these parameters in high-quality unit-selection synthesis has to be developed (Schröder, 2009).

Thus, speech synthesis is a useful tool for studying emotions and speech, because it affords good controllability of the timing and prosodic cues. As mentioned, research purely concerning the effects of the lexical emotional content of speech on the human emotional system is rare. Thus, the present thesis concentrated on studying the effects of the lexical content of messages, instead of the prosody of the speech. In addition, in interfaces and applications that utilize speech as an output method, speech synthesis has some advantages over the recordings of human speech. For example, if the prerecorded voice prompts are used, the application has to include all possible prompts, and thus, storage requirements are huge. The memory requirements of speech synthesizers are not so large.



4 Human-Computer Interaction

The HCI research area focuses on studying and designing the interactive elements between people and computers. A basic idea behind the field of HCI has been to make computer use more natural, effective, and enjoyable. One possible approach toward this goal is to design interfaces so that they mimic the ways in which humans communicate with each other, such as by utilizing speech as an interaction method between humans and computers. On the other hand, as studies in last few decades have shown, emotions play an important role in all aspects of human life, such as in communication, well-being, and human rational behavior. Thus, emotions clearly affect human behavior in HCI too. In this section, first, a brief introduction to two specific research areas that are especially important to emotions and HCI is given. These are affective computing and user-experience research. Then, a relevant paradigm for affective HCI (i.e. computers are social actors (CASA)) is introduced. Finally, the role of anthropomorphism in HCI is discussed.

4.1 Emotions in Human-Computer Interaction

According to Mayer and Salovey (1997, p. 10), emotional intelligence includes the abilities to perceive, appraise, and express emotions. Clearly, these emotional abilities have effects on the quality of interpersonal communication and interaction. Thus, the emotional intelligence of a computer, at least on some level, would also make the interaction between humans and computers more pleasant and effective. As Picard (2003) has mentioned, of course, not every machine needs emotional abilities, but in many situations, it would be beneficial. Two research areas within HCI that are related to affective issues are affective computing and user-experience research.

Affective computing is a relatively new research field, which aims to integrate emotions into HCIs. As already mentioned in the introduction section, Rosalind Picard published a pioneering book, *Affective Computing*, in 1997, in which she defined affective computing broadly as "computing that relates to, arises from, or deliberately influences emotions." To be emotionally intelligent, a computer needs the knowledge of its users' emotions, and the capacity to express emotionally meaningful cues that have desirable effects on those users (Picard, Vyzas, & Healey, 2001; Ochs, Niewiadomski, Pelachaud, & Sadek, 2005). Thus, the ultimate goal of affective computing is to design and create computer systems that can recognize and interpret users' emotions, and respond appropriately to these emotions (Surakka & Vanhala, 2011).

Machines can acquire information about human emotions from different signals. These include expressions in the voice, face, body movements, or text, and changes in physiology (Cowie, Douglas-Cowie, Tsapatsoulis Votsis, Kollias, Fellenz, & Taylor, 2001; Lisetti & Nasoz, 2004; Calvo & D'Mello, 2010). In some research, computer algorithms have recognized emotions from speech or facial expressions at a level that is as good as or even slightly better than humans can manage (see Picard et al., 2001). Partala, Surakka, and Vanhala (2006) explored how accurately a computer could estimate participants' emotional experiences from their facial expressions (activity of the zygomaticus major and corrugator supecilii) when the participants viewed emotionally negative, neutral, and positive pictures and videos. The results showed that the best models could estimate negative and positive ratings with an accuracy of over 70 and 80% for pictures and videos, respectively. As emotions activate multiple physiological and behavioral response systems, it would be advantageous to acquire information from different modalities. There is proof that recognition rates are higher in systems that gather information through multiple modalities, as compared to unimodal systems. For example, Castellano, Kessous, and Caridakis (2008) trained and tested a model that aimed to detect eight emotions integrating information from the facial expressions, body movements, gestures, and speech. Recognition rates were more than 10% higher for the multimodal classifier in comparison with unimodal systems. The results of the recognition studies sound promising, but, on the other hand, the high recognition rates are acquired from a small set of rather exaggerated, portrayed expressions or singly occurring facial actions, and more research is needed before affect detection is functional in real time with naturally occurring expressions (Picard et al., 2001; Calvo & Mello, 2010). All in all, in spite of the progress made in emotion detection, it is still a very challenging area of research that is in its infancy.

Machines do not have emotions, but it is possible to program them to manifest emotion-like expressions. Several possible modalities can be used

to convey affective information (e.g. computer agents' facial expressions, body movements, gestures, and speech). As presented in the preceding section, people can recognize emotions from the prosodic cues of synthetic voices quite well, even if the voice sounds clearly artificial and machine-generated. There is also evidence that people recognize emotions well from the facial expressions of a computer-generated face (e.g. Bartneck, 2001; Ku, Jang, Kim, Kim, Park, Lee, Kim, Kim, & Kim, 2005; Dyck, Winbeck, Leiberg, Chen, Gur, & Mathiak, 2008). However, how people respond to synthetic emotional expressions is for the most part not yet understood. Beale and Creed (2009) have concluded in their review that in many studies, the use of emotion had little positive impact on the interaction, but equally so, in many studies, the use of emotions enhanced the interaction. For example, Brave, Nass, and Hutchinson (2005) found that people rated a computer agent that expressed other-oriented, empathic emotion as more likeable and trustworthy than a computer agent that expressed self-oriented emotion. In addition, an agent's empathic responses can reduce a user's stress and frustration in computerized game sessions (Partala & Surakka, 2004; Prendinger, Mori, & Ishizuka, 2005; Hone, 2006), and can have a positive impact on learner interest in a computerized learning environment (Kim, Baylor, & Shen, 2007).

Almost all existing research has concentrated on user attitudes, perceptions, or behavior, but still less is known about how these synthesized emotional expressions induce emotions in people. Seeing or hearing the emotional expressions of other people can affect people's own emotions; that is, watching the facial expressions of sadness and happiness causes corresponding facial expressions and affective experiences in an observer, even in controlled laboratory settings, or when the expressions are low in intensity (Hsee, Hatfield, Carlson, & Chemtob, 1990; Surakka & Hietanen, 1998; Hess & Blairy, 2001). Further, hearing the vocal expressions of anger and contentment can cause changes in a listener's physiology and experience (Hietanen et al., 1998). These kinds of findings suggest that emotions are contagious, so that people have a "tendency to mimic the verbal, physiological, and/or behavioural aspects of another person's emotional experience/expression, and thus to experience/express the same emotions oneself" (Hsee et al., 1990, p. 328). Primarily, the work on emotional contagions has concerned exploring the effects of the nonverbal expressions of emotions, but there is some evidence that the verbal content of speech may have similar effects. Hsee, Hatfield, and Chemtob (1992) examined how observing another person's facial expressions and spoken messages that are conflicting affects the observer's appraisals of the other person's emotional state and their own emotional state. The participants watched a video in which a stimulus person actually experienced and expressed a happy or a sad emotion and at the same time reported being either happy or sad. The messages were

spoken by a flat, machine-like, and unemotional voice. The results showed that the appraisals of the stimulus person's emotions were primarily affected by the verbal message. Participants' own emotions were influenced equally by the verbal message and stimulus person's facial expressions.

The question of the emotional contagion in HCIs remains largely unanswered. It is not yet known if the emotions expressed by a computer can be contagious; that is, is it possible to evoke exactly the same emotion in a person that a computer is expressing? Prendinger, Becker, and Ishizuka (2006) found that the negative empathic behavior of an agent induced negatively valenced emotion (measured through facial EMGs). Thus, they suggested that there was at least a certain degree of reciprocity between an agent's expression and a user's response. Weyers, Mühlberg, Hefele, and Pauli (2006) also measured facial EMG responses to synthesized expressions of emotions. They found that viewing happy facial expressions of an avatar evoked congruent facial muscular reactions in viewers (i.e. increased zygomaticus major and decreased corrugator supercilii activation). It is noteworthy that angry facial expressions elicited no significant EMG activation. In the study by Ku et al. (2005), avatars' emotional facial expressions were well recognized, but the participants were not influenced by the facial expressions to the same degree.

There is also some evidence that emotionally loaded spoken messages can regulate HCIs. One study showed that positive emotional feedback given by the monotonous voice of a speech synthesizer during computerized problem-solving tasks facilitated cognitive performance and recovery from autonomic arousal, even when the feedback did not relate to actual performance, but was fully random (Aula & Surakka, 2002). In other research, participants were exposed to mouse delays during computerized problem-solving tasks (Partala & Surakka, 2004). The results showed that problem-solving performance was significantly better and smiling activity higher when participants received positive affective intervention via a speech synthesizer during mouse delay than when they received no intervention.

Although the definition of affective computing is broad, the field mainly deals with work that aims to examine and develop technologies that can sense and recognize the emotional states of users, that aims to construct and develop the emotional expressivity of computers, and to develop systems that can adapt their own functioning to meet with their users' emotional state. Thus, as mentioned above, it is largely unclear how people respond to the affective expressions of computers. On this account, it has been claimed that affective computing takes a computer perspective to emotions, thus placing little emphasis on the emotions of users that arise when people interact with technology (Hassenzahl & Tractinsky,

2006). In addition to affective computing, there is another approach to examining emotions in HCIs; that is, user-experience (UX) research. Traditionally, HCI research has concentrated on studying and developing the usability of computer systems. That means that a basic principle in designing interfaces and systems has been to build systems whose use is as easy and as effective as possible. Traditionally, usability has focused on the cognitive aspects of technology use. Recently, it has been realized that users' perceptions and experiences during interaction are at least as important as the other aspects of use. There is no unanimously accepted definition of the concept of UX, but it refers to a wide variety of experiences that emerge during human–technology interaction (Law, Roto, Hassenzahl, Vermeeren, & Kort, 2009). Affective experiences are recognized as a central part of these experiences. By studying human experiences, UX research focuses on the consequences that the technology has on humans. Thus, it has been proposed that UX research takes a human perspective in terms of emotions (Hassenzahl & Tractinsky, 2006).

The present thesis somewhat relates to both research areas. On the one hand, it is linked to affective computing, particularly to the research direction, which aims to design and create the emotional expressiveness of computer systems. The first studies on emotional expressiveness showed that people could recognize emotions quite well from the synthesized expressions of emotions (e.g. Cahn, 1990; Murray & Arnott, 1995; Bartneck, 2001; Ku et al., 2005; Dyck et al., 2008). Recently, the research has also started to examine how people respond to these emotional expressions. This thesis aims to contribute to this debate, by examining if it is possible to evoke emotion-related responses in people through synthesized, spoken verbal emotional expressions. On the other hand, specifically looking at people's emotion-related responses also connects the present study to UX research.

4.2 Computers Are Social Actors

A relevant research frame of reference for affective HCIs is the *computers are social actors'* paradigm (Reeves & Nass, 1996; Surakka & Vanhala, 2011). The paradigm suggests that people treat computers and other media as they treat other people; that is, socially. The first findings on this area came from the experiments carried out at Stanford University by Clifford Nass and colleagues. Their idea was to conduct studies, which are similar to classical social science experiments on human–human interaction, with the exception that one of the human actors in the social science studies was replaced with a computer actor. These replicated studies have shown that people are polite to computers (Nass, Moon, & Carney, 1999), that people are affected by flattery coming from a computer in a similar manner to flattery coming from another person (Fogg & Nass, 1997), and that people regard computers as teammates (Nass, Fogg, & Moon, 1996).

Later work by other researchers has extended the research. For example, Ferdig and Mishra (2004) found that people got angry and punished a computer when they felt that the computer had treated them unfairly. It has also been found that positive feedback from a computer affects participants' motivation and affect positively (Mishra, 2006). In addition, Vanhala, Surakka, Siirtola, Räihä, Morel, and Ach (2010) showed that by imitating human social cues such as adjusting the virtual proximity of the computer character, it is possible to elicit different subjective and physiological responses in people.

Importantly, many studies have shown that people respond similarly to speech regardless of whether it comes from a human or a computer (Nass & Brave, 2005). For example, Nass et al. (1997) found that the male and female voice of a computer evoked gender-based stereotypic responses toward computers, with people rating the evaluation from a male-voiced computer as more valid than evaluation from a female-voiced computer. In addition, participants perceived a female-voiced computer as more informative about feminine topics, and a male-voiced computer as more informative about the topics that are typically regarded as masculine. Further, Nass and Lee (2001) showed that people consider different voices as separate individuals, even if the voices were clearly nonhuman, and the same computer delivered the voices. Participants were also attracted to the voice that was similar to their own personality.

These findings do not reflect that people think that computers have feelings, gender, or personalities, but that people react to computers as if they had them (Reeves & Nass, 1996; Ferdig & Mishra, 2004). Nass and Moon (2000) have suggested that people's tendency to apply social rules and expectations to computers results from mindlessness. This means that social responses toward machines are automatic, unconscious, and natural. The results were independent from the age of the research participants or their experience in computer use. In addition, the participants were well aware of the fact that they were communicating with a computer, and although they had argued that they would never respond socially toward a computer, they still responded in just such a social manner. The mindlessness responses arise from the fact that the human brain has evolved for social interaction with other humans; it is not specialized for interaction with new media. Thus, if there are enough cues that suggest humanness (i.e. social cues, written words, speech, etc.), people unconsciously and automatically categorize computers as social actors, and respond to them accordingly.

4.3 ANTHROPOMORPHISM

In 1944, Heider and Simmel studied how people described a short, silent animation video in which a small triangle, a small circle, and a large

square moved around each other. In addition, there was a rectangle with a section that could be opened and closed as a door. The task was to watch the film and discuss what happened in it. Almost all of the people described the movements of the objects as social interactions. The moving objects also elicited the interpretations of personality, gender, and the emotions of the objects. Thus, simple, abstract, geometric shapes activated anthropomorphic interpretations in people.

Anthropomorphism means the attribution of human characteristics and behavior to nonhuman things, including animals, spiritual deities, and electronic devices (Guthrie, 1993; Epley, Waytz, & Cacioppo, 2007). Epley and colleagues (2007; 2008) developed and tested a three-factor theory of anthropomorphism to explain why people are likely to anthropomorphize nonhuman agents. First, people have much more detailed knowledge about humans, and this knowledge is acquired earlier than knowledge about nonhuman agents. This rich knowledge may activate automatically when people judge, predict, and try to understand nonhuman, unknown agents and their behavior. Secondly, there are two motivational mechanisms that are important determinants of anthropomorphism: effectance and sociality. Attributing human characteristics to nonhuman agents may fulfill people's basic need to understand, control, and predict an otherwise uncontrollable world. In addition, people have a fundamental need and desire for social connection with other people. The anthropomorphization of nonhuman agents enables human-like connections with them. For example, Epley et al. (2008) found that lonely people were more likely to anthropomorphize their pets than participants who had more social connections.

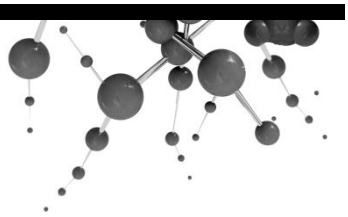
In the field of HCI, the concept of anthropomorphism has also been used in another sense; that is, it can refer to the similarity between a human and a computer interface (e.g. Schaumburg, 2001; Nowak & Rauh, 2008). More specifically, the similarity can refer to the human-like characteristics of an interface, such as its face and voice, or human-like capabilities and behavior. Further, anthropomorphism is not a dichotomy, but the appearance, voice, or behavior of an interface agent can vary along a dimension from less to more human-like. These two meanings of anthropomorphism may be interconnected, at least up to a point. For example, Epley et al. (2007) have suggested that the more the observable features of a nonhuman agent resemble humans, the more people are likely to anthropomorphize the agent.

There is some research on how humanizing interfaces can affect computer users. One area has focused on examining the effects of human-like features on a user's social responses. Reeves and Nass (1996) have shown that people respond in a social manner to computers, even though the interface is a simple text-based system without anthropomorphic features

such as speech or animation. However, Nass and Moon (2000) suggested that it might be possible that this tendency is even stronger when the computer is more human-like. Sproull, Subramani, Kiesler, Walker, and Waters (1996) studied people's responses to a computer-based career-counseling system. The participants used either a talking-face or a text-display interface. The face looked like a real human, but a formant synthesizer produced the voice, and thus it sounded artificial. The results showed that the participants responded more socially to the face–voice interface, but especially women rated the text-display more positively than the face–voice display. In another study, participants interacted with the same talking face, a text interface, a voice interface, or with a human partner in a cooperative prisoner's dilemma game (Kiesler, Sproull, & Waters, 1996). Face-to-face communication with a real human received the highest sociality evaluation rating, and it evoked the strongest feeling of partnership. The participants had stronger feelings of partnership in the face–voice condition than in the voice or text conditions, but in line with Sproull et al. (1996), the face–voice condition (i.e. the most human-like computer condition) was rated as the least attractive and warm of all the partners. It was later argued that the negative ratings of the talking-face interface may have been due to the mismatch between the real-looking face and the mechanistic voice (Gong, 2008). That is because several studies have shown the importance of consistency between the voice, look, and behavior of an interface for it to be comfortable for the interaction partners (Nass & Brave, 2005). More recently, Gong (2008) found that the more human-like a computer agent was, the more it received social responses. The research was limited to facial images to avoid possible inconsistencies between different modalities.

Previous studies have also explored how people perceive the different levels of human-likeness. Power, Wills, and Hall (2002) found that people experienced the most realistic looking interface agents as scary, but that they also rated them as more intelligent than abstract characters. Instead, abstract characters were rated as being more friendly and pleasant, but as less interesting and boring. Further, in some research, less human-like avatars have been perceived as more credible and likeable than more human-like avatars (Nowak, 2004), and in other work, more human-like avatars have been rated as more credible and attractive (Nowak & Rauh, 2005). Overall, Dehn and van Mulken (2000) concluded in their review that the results of the effects of human-likeness are inconsistent. These findings are somehow contradictory to the suggestion of the "uncanny valley" by Mori (1970). According to the uncanny valley hypothesis, people's responses to nonhuman agents are straightforwardly more positive the more the agent resembles a human up to a certain point, at which the reaction becomes negative. At this point, the agent looks highly, but not fully human-like.

Taken together, people do seem to respond, for example, more socially to computers the more the computer resembles a human, but overall, the effects of human-likeness are not clear. Clearly, more research is needed to study the effects of human-like qualities. Moreover, Dehn and van Mulken (2000) suggested that the approach has to be more detailed. One study cannot solve the whole question, but it is necessary to study in what situations what kinds of features promote the interaction. As computer interfaces that express emotions are becoming increasingly popular, it is important to know how the human-likeness of a computer that expresses emotions affects the emotional responses of humans. There is evidence that the facial emotional expressions of machines are as convincing as the expressions of natural human faces (Bartneck, 2001). On the other hand, Weyers et al. (2006) found that dynamic, and therefore more natural facial emotional expressions of an avatar evoked stronger emotion-specific facial-muscle reactions in viewers than static expressions. All in all, the question of the relationship between the anthropomorphism of a computer interface and emotions has remained mainly unexplored and the present thesis aims to contribute to this question.



5 Experiments

The aim of the work in this dissertation is to explore how people respond to synthetically produced lexical expressions of emotions. A common aim herein was to investigate if it was possible to evoke emotion-related changes in people through the lexical emotional content of a synthesized voice. The second common aim for the work reported in Publications I, II, III, and V was to investigate if the naturalness or human-likeness of a synthetic voice affected human responses. Additionally, one aim of the study, reported in Publication V, was to explore if the emotional content of the spoken messages affected the ratings of voice quality.

5.1 Publication I: Subjective and Physiological Responses to Emotional Content of Synthesized Speech

The aim of the first paper was to examine the effects of synthesized speech with emotional content on participants' facial EMG activity and subjective experiences. The negative, neutral, and positive sentences that described either the computer's emotional state (e.g. I am angry) or appraised the participant's emotional state (e.g. You look happy) were produced by two different speech-synthesis techniques: synthesis based on the microphonemic method (Mikropuhe©) and diphone synthesis (Suopuhe). The sentences were produced using the male voice of both synthesizers. The length of each sentence was approximately the same: 2.2 seconds. The F0, speech rate, and volume were the same for both synthesizers. The F0 variation was set as flat as possible to keep the prosody of the voices as neutral as possible. Thus, the emotions were expressed only by the emotional meaning of the sentences, and it was also possible to compare if the synthesizers affected the physiological and subjective responses differently. Fifteen participants (10 females) listened to the stimuli while

their EMG activity over the corrugator supercilii (frowning activity) and zygomaticus major (smiling activity) areas were measured. After the listening phase, the participants heard the sentences again, and rated their emotional experiences on the valence and arousal dimensions.

The results showed that the sentences produced by both speech synthesizers evoked significant differences in the participant's ratings for different stimulus categories. The positive sentences generated by both synthesizers evoked significantly more positive ratings for valence than the neutral or negative sentences. The valence ratings between the negative and neutral sentences were not significantly different for either of the synthesizers. The negative sentences produced by Suopuhe evoked higher ratings of arousal than the neutral or positive sentences, while the arousal ratings were not different for the different emotion categories produced by Mikropuhe©. Further, the sentences that appraised participants' emotional state were rated as significantly more arousing than the sentences that described the computer's emotional state. The facial EMG data was analyzed in two time windows: during the sentence and during the three-second time period following the stimulus offset. The activity of the corrugator supercilii muscle site decreased significantly more during and after the positive sentences than during and after the neutral sentences generated by Suopuhe, the more human-like speech synthesizer. The activity of the zygomaticus major increased during and after all stimulus categories, but the categories had no statistically significant effect on the activity of this muscle site.

In sum, the results showed that even the emotional content alone can elicit the changes in the ratings of emotional valence, arousal, and facial-muscle reactions. However, only the more natural voice caused significant changes in the ratings of arousal and emotion-related facial-muscle responses. Thus, it seems that increasing the naturalness of the voice can enhance the effects of the spoken messages.

5.2 Publication II: Emotions, Anthropomorphism of Speech Synthesis, and Psychophysiology

The second study aimed to investigate the effects of synthesized speech with lexical emotional content on participants' pupillary responses and subjective experiences. Twenty-six participants (15 females) took part in the experiment. The experimental design was the same as in the study reported in Publication I. The pupil size was measured unobtrusively and remotely by using a floor-mounted eye tracker.

The results showed that the emotional sentences produced by both speech synthesizers evoked significantly different ratings of emotional valence. The positive sentences resulted in more positive ratings of valence than

the negative or neutral sentences, and the negative sentences resulted in more negative ratings of valence than the neutral sentences. Instead, the analysis of arousal ratings and pupil size revealed that the sentences generated by Suopuhe, the more human-like synthesizer, evoked more variation both in the subjects' physiology and subjective experiences than the sentences generated by Mikropuhe©. More specifically, there was a significant curvilinear relationship between the negative, neutral, and positive sentences in terms of pupil size and arousal ratings when Suopuhe produced the sentences. Further, pupils dilated significantly more after the negative and the positive sentences than after the neutral sentences when Suopuhe generated the sentences. When the sentences were generated by Mikropuhe©, neither the pupil size nor arousal ratings were different between the stimulus categories.

The results of this study supported the findings of the first study. Both synthesizers were able to elicit significant changes in the ratings of valence, but only the more human-like voice evoked significant emotion-related changes in the ratings of arousal and pupil responses. Thus, again, the more human-like voice enhanced the emotional responding of the participants to messages with emotional content.

5.3 Publication III: Subjective Responses to Synthesised Speech with Lexical Emotional Content: the Effect of the Naturalness of the Synthetic Voice

The purpose of the third study was to investigate further how the verbal emotional content and the degree of the naturalness of synthesized speech would affect the participants' ratings of their emotions. Secondly, it aimed to examine how the naturalness of the synthesized voice would affect the voice-quality ratings. In contrast to the studies reported in Publications I and II, this experiment included all the techniques that have been dominant in creating synthesized speech: formant synthesis (Saarni, 2010) and two concatenative synthesis techniques, diphone synthesis (Suopuhe) and unit-selection synthesis (Bitlips). The stimuli consisted of words that were selected (and translated into Finnish) from the standardized set of emotional words, from the ANEW (Bradley & Lang, 1999). Five of the stimulus words were highly negative and arousing, five highly positive and arousing, and five were neutral. The words were generated by the male voice of the speech synthesizers. The pitch, loudness, and speech rate of the synthesizers were adjusted to the same level. In addition, the words were produced so that the F0 variation was set as flat as possible to keep the prosody of the voices as neutral as possible.

The experiment consisted of two phases: a listening and a rating phase. In the listening phase, 24 participants (12 females) heard each stimulus and indicated whether they understood the word by pressing one of two

buttons on a response box. After the listening phase, participants heard all the words again and rated them on dimensions of valence, arousal, and approachability. Finally, the participants heard the neutral words again and rated the pleasantness, naturalness, and clearness of the voice. The clearness refers to the clarity of the synthesizer's articulation in this work.

The results showed that the words produced by the two more human-like speech-synthesis systems—diphone and unit-selection synthesis—were more intelligible than the words produced by the machine-sounding formant synthesis. A higher number of the words produced by formant synthesis were not understood in the listening phase as compared to the words that were generated by the concatenative synthesizers. The analysis of the emotional ratings showed that even though the affective words produced by all speech synthesizers evoked significantly different ratings of emotional valence, arousal, and approachability, emotionally negative and positive words produced by the concatenative synthesizers, the more human-like synthesizers resulted in more pronounced ratings of valence and approachability than the words produced by the formant synthesizer. The voice-quality ratings of the neutral words showed that participants rated the voice of the unit-selection synthesis as the most natural. Otherwise, the participants rated the voices of both concatenative synthesizers as more pleasant and clear than the voice of the formant synthesizer. An additional analysis that was not part of the original publication showed that there was a statistically significant correlation between the ratings of naturalness and pleasantness, $r = 0.65$, $p < 0.001$ ($r^2 = 0.42$). Thus, this explains 42% of the variation between naturalness and pleasantness.

The above results showed that even a very machine-like voice can be used to communicate emotional messages. The stimuli evoked significant changes between the stimulus categories in the ratings of valence, arousal, and approachability, even though the stimuli consisted of very short words. However, also in this study, the more human-like voices caused more enhanced ratings of emotions. That is, the ratings for the experienced valence and approachability were significantly stronger when the more human-like voices produced the words than when the less human-like voice produced the words.

5.4 Publication IV: Heart Rate Responses to Synthesized Affective Spoken Words

This study investigated heart-rate responses and emotion-related ratings to synthetically produced spoken words. While 20 participants listened to the emotionally negative, neutral, and positive words that were also used in study III, their heart-rate responses were measured unobtrusively by using a regular-looking office chair, the EMFi chair. The EMFi chair

measures continuous ballistocardiographic heart-rate responses with electromechanical film (EMFi) sensors that are embedded in its seat, backrest, and armrests.

The results revealed that the ratings for the words were in accordance with their lexical valence. In addition, the positive words were rated as more approachable than the neutral or negative words, and the neutral words were rated as more approachable than the negative words. The ratings for arousal were higher for the negative than the neutral or positive words. The difference between the positive and neutral words was not significant.

The results showed further that the synthesis technique did not affect the heart-rate responses. However, heart-rate deceleration was greatest for the negative words during a five-second period after the stimulus offset. Further, the second-by-second analysis showed that at the fifth second from the stimulus offset, the heart rate was more decelerated after the negative words than after the positive words. Thus, heart-rate deceleration was the strongest and most prolonged for the negative stimuli. The results of this study suggest that brief synthetically produced spoken words with emotional content can evoke similar decelerated heart-rate response patterns as found earlier, only with longer stimuli.

5.5 Publication V: The Effects of Emotionally Worded Synthesized Speech on the Ratings of Emotions and Voice Quality

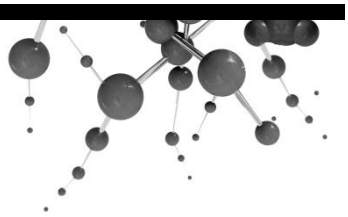
The aim of this study was to investigate how synthetic verbal stimulation with emotional content might affect the perceiver. First, similarly to the earlier studies, it aimed to examine how the verbal content of synthetic messages and the level of voice naturalness affected participants' emotional responding. Second, it assessed whether the ratings for voice quality were affected by the emotional content of the spoken messages.

The stimuli consisted of twelve different sentences with emotional content: angry, happy, and neutral. We selected happy and angry words from the list of causative emotions by Johnson-Laird and Oatley (1989), and translated them into Finnish. Neutral sentences were constructed so that the sentences in different categories matched in terms of word length. The sentences were generated by a male voice from the unit-selection speech synthesizer (Bitlips) with monotonous and neutral prosody. By a monotonous tone of voice, we mean that the variation in F0 was set as flat as possible. By a neutral tone of voice, we mean that the variation in F0 was not manipulated; that is, the variation in F0 was not flattened, nor were any emotional cues added. The other voice parameters (i.e. speech rate, loudness, and pitch) were the same for both speaking styles. The purpose of this F0 manipulation was to study whether emotional effects

could be enhanced by increasing the prosodic variation of the speech. Twenty-eight participants (21 females) listened to the sentences while their facial EMG activity over the corrugator supercilii was measured. After the listening phase, the participants gave their ratings for emotions and voice quality.

The results showed that the content of the sentences affected the ratings of experienced valence. The ratings of valence were different for all emotion categories. The results also revealed that the tone of the voice affected the ratings of arousal. The sentences generated by the neutral tone of voice evoked more emotion-relevant ratings of arousal than the sentences generated with the monotonous prosody. As expected, the participants rated the prosodic tone of voice as more pleasant, natural, and clear than the monotonous voice. An additional analysis that was not part of the original publication showed that there was a statistically significant correlation between the ratings of naturalness and pleasantness, $r = 0.72$, $p < 0.001$ ($r^2 = 0.52$). Thus, this explains 52% of the variation between naturalness and pleasantness. Interestingly, however, the verbal content of the sentences had an effect on the ratings of voice quality. When the content of the sentence was positive, the quality of voice was rated more positively than when the content was neutral or negative. In this study, the EMG responses did not seem to reflect emotional reactions, but cognitive processing (concentration) instead. Frowning activity decreased significantly from the pre-stimulus baseline after the neutral sentences, whereas the activity of the corrugator muscle did not differ significantly from the pre-stimulus baseline after the negative and positive sentences. Thus, the EMG results suggest that participants concentrated less on the processing of the neutral sentences than on the processing of the emotionally loaded sentences.

In sum, in line with earlier studies, the ratings of emotions were significantly affected by the emotional content of the sentences. In addition, the ratings for arousal were more emotion-related when the sentences were produced by the prosodic (i.e. more natural) tone of voice than when the sentences were produced by the monotonous tone of voice. However, although the speaking style had some effects on the ratings for arousal, overall, the speaking style had quite a small effect on the ratings of emotions. This may arise from the fact that the voice itself was very natural. More interestingly, the results showed that the content of the sentences affected the ratings of voice quality. Thus, the results suggest that positive content could be used to enhance positive feelings about the voices.



6 Discussion

First, the results presented in this thesis showed that the emotional messages produced by synthesized speech had significant effects on people's physiology and the ratings of their emotional experiences. This was true even though the emotional messages were delivered without any interaction context between the human and the computer. In other words, passive listening to verbal emotional material automatically induced changes both on the subjective and physiological levels. The ratings of experienced valence, arousal, and approachability were mainly different for the different emotion categories. In addition, the results reported in Publications I, II, and IV also showed that the facial EMG, pupil, and heart-rate responses were modulated by the emotional content of the spoken message.

All the studies in this thesis utilized the subjective ratings of emotions to gain information regarding participants' emotions. Some also utilized physiological measurements. Based on the subjective ratings, it is possible to say that the emotional stimuli evoked the changes at least in the cognitive ratings of emotions. However, it is possible that the stimuli also evoked changes in terms of spontaneous reaction. This is supported by the findings of studies I, II, and IV. The results of these studies showed that the lexical emotional content of the synthesized speech had an effect on the facial EMG, pupil, and heart-rate responses. Thus, in line with the previous research that has found that the facial expressions of computer characters can elicit significant emotion-related physiological responses (Weyers et al., 2006), we also found that non-embodied synthesized emotional expressions can evoke emotions in people.

Second, the results provided evidence that the features of the voice are relevant when we want to evoke emotional reactions through synthetically

produced spoken messages. The results of Publication I showed that the corrugator EMG activity attenuated significantly more during and after the positive sentences than during and after the neutral sentences, but these differences were significant only when the more human-like speech synthesizer generated the sentences. The second study revealed that the sentences produced by a more human-like synthesizer caused more variation in both the pupil responses and subjective experiences than the sentences produced by a less human-like synthesizer. In study III, we found that the ratings for experienced valence and approachability were stronger when the sentences were generated by the more human-like voices as compared to the more machine-like voice. Finally, study V indicated that the averages of the ratings for arousal were more emotion-related when the neutral tone of voice produced the sentences than when the monotonous voice produced the sentences.

First of all, the findings of this study have highlighted the importance of language in human communication. Even though the spoken stimuli were generated by the monotonous voices of speech synthesizers, the stimuli activated the human emotion system. Spoken words (i.e. the language lexicon) provide a rich source for people to express their feelings, reactions, and judgments (Van Lancker Sidtis, 2008, p. 200). Thus, spoken language has an important and central role in interpersonal communication and in the building of social relationships (Pinker, 1995). The reason why even artificial emotional messages may be particularly significant in evoking emotional processes may relate to the fact that hearing has some benefits as compared to vision or to other senses. For example, hearing is mostly independent of the position of the head and ears, while the field of vision is highly dependent on the direction of the eyes. Therefore, the hearing system has served as a primary system for receiving warnings from others, and the hearing of verbal utterances may have had a significant meaning for survival in many situations (Scharf, 1998). For these reasons, people may be highly responsive to spoken emotional messages. Speech has had an important role in human evolution, and it is one of the most important means of human communication. Thus, speech seems to offer a good medium through which to provide emotional messages that aim, for example, to warn or arouse people in the context of HCIs.

Previous research concerning emotions and speech has concentrated mainly on emotional prosody. It has been found that people can recognize emotions fairly well from the prosodic cues (Banse & Scherer, 1996) and the vocal expressions of emotions can evoke emotion-related psychophysiological responses (e.g. Hietanen et al., 1998). However, in the light of the current results, it seems that the emotion-related prosody of speech is not necessary to be able to communicate emotional messages. Of course, prosodic speech cues have a central and important role in conveying information about the emotional state of a speaker, and

probably adding prosodic cues would have further increased the power of the messages of the current study. However, the content itself seemed to be powerful enough to evoke emotion-related responses in the listeners.

Nass and Brave (2005) have conducted a series of studies regarding the effects of speech in HCIs, and they have found that people respond in the same way to speech whether it comes from another human or from a machine. They suggested that the responses are similar because "people do not have separate parts of the brain for human speech and technology-generated speech" (p. 183). Even though our results showed that it is possible to evoke emotions in listeners through technology-generated speech, our findings also showed that the naturalness of the voice matters when we communicate emotional messages from a computer to a human. The naturalness of the voice affected how strong people's responses to the voice were. Further, the results of Publication III revealed that the participants rated the pleasantness of the voice as higher when the more human-like voice produced the stimuli than when the less human-like voice produced the stimuli. Thus, the findings of this thesis support the importance of anthropomorphism, at least in audio components, and at least when communicating emotional messages from a computer to a human.

The earlier findings about the effects of anthropomorphism have been somewhat contradictory, and they have mainly concentrated on examining the effects of visual appearance. For example, Nowak (2004) and Nowak and Rauh (2005; 2008) looked at how the anthropomorphism of avatars affects people's perception. In some research, less anthropomorphic avatars were liked more than avatars that were more anthropomorphic, while in some work, it was found that the participants rated avatars that were more anthropomorphic more positively than less anthropomorphic avatars. Meta-analysis by Yee, Bailenson, and Rickertsen (2007) has shown that the interaction with interfaces with an embodied agent produces more positive social interaction than interaction without agents. The human-likeness of a representation also had a significant but minor effect on the interaction. Thus, they concluded that the presence of a human-like character is much more important for the quality of interaction than the human-likeness of an agent. One widely known phenomenon related to the human-like appearance of a robot or other artificial human-like object and the emotional response of a human is the uncanny valley hypothesis. According to the hypothesis that was first proposed by a Japanese roboticist Masahiro Mori in 1970, the more a robot looks and moves like a real human, the more positive the viewer's reactions are, up to a certain point at which the reaction becomes negative. In this valley-shaped dip in the proposed graph, a robot looks highly but not fully human-like, and it is rated as more eerie and unpleasant than familiar and likeable. Mori's original hypothesis was based on informal

observations and predictions and not on empirical tests (Seyama & Nagayama, 2007; Misselhorn, 2009). Much recent research exploring the phenomenon with static pictures and video clips while manipulating the human-likeness of characters have proposed that the relationship between human-likeness and people's reactions seems to be more complex than the graphical representation of Mori's theory. For example, Tinwell and colleagues (2009, 2011) have suggested that there might be more than one valley when the ratings of perceived familiarity are plotted against human-likeness. Further, they suggested that the uncanny valley could be replaced by the uncanny wall concept, because it seems that the uncanny valley cannot be traversed. That is, even very human-like, realistic characters were judged as being less familiar than a real human being. In addition, Ho, MacDorman, and Dwi Pramono (2008) have found that an unpleasant reaction is not necessarily the result of the degree of human-likeness, but that it may result from the inconsistency between more and less human-looking elements. In line with this finding, Seyama and Nagayama (2007) proposed that the uncanny valley was related to the abnormal features of facial images. In their study, the participants rated almost human-like facial images as negative only when there was an abnormal feature in the image's face (i.e. enlarged eyes with a photorealistic facial texture).

The present results give some support to the findings that the relationship between people's responses and human-likeness is not necessarily straightforward. In Publication III, we found that the voice of unit-selection synthesis was rated as the most natural, as expected. Both concatenative synthesizers (i.e. unit-selection and diphone synthesis) were rated as more pleasant and clear than the robot-like formant synthesizer, but the ratings of pleasantness and clarity did not differ between diphone synthesis and unit selection. The correlation analysis revealed a significant connection between the ratings of naturalness and pleasantness. That is, the higher the participant rated the naturalness of the voice, the more pleasant she/he rated the voice. In Publication V, the naturalness of the voice was increased by adding prosodic variation to speech produced by the unit-selection synthesizer. The results showed that the prosodic tone of voice was rated as more pleasant, natural, and clear than the monotonous voice. In addition, in line with the findings of Publication III, there was a significant correlation between the ratings of naturalness and pleasantness. Thus, increasing the naturalness of the voice seems to have a positive impact on the ratings of pleasantness, and increased naturalness did not cause any dip in the ratings of pleasantness.

However, in Publication III, the ratings of the valence and approachability of the positive words were enhanced when the words were produced via diphone synthesis, and the ratings of the valence of the negative words were enhanced when the words were produced via unit-selection

synthesis as compared to formant synthesis. Thus, the most natural voice had an impact especially for the ratings of the negative words but not for the ratings of the positive words. One possible reason for this is probably related to the inconsistency between more and less human-like elements. The voice of the unit-selection synthesizer sounded the most natural, but on the other hand, all the voices were very monotonous. Waaramaa-Mäki-Kulmala (2009, p. 74) noted that people usually interpreted human speech without any affective cues as unfriendly. In addition, Bertels et al. (2009) ascertained that people experienced negative words uttered in a neutral tone of human voice as more threatening than the same words uttered in an emotionally congruent tone of human voice. They suggested that it is possible that the neutral tone of voice evokes an impression of coldness. Thus, it is possible that in our study, the human-like voice with a monotonous speaking style was experienced as cold, and thus it enhanced the effects of the negative words. In sum, even though some of the findings of this thesis support that increasing the naturalness of the voice evokes more positive perceptions in people, they also suggest that the relationship between anthropomorphism of the voice and of human behavior is not necessarily straightforward. The consistency between the different elements seems to also affect the experiences of voice as previously suggested for visual stimuli.

Emotionally relevant corrugator EMG responses to affective sentences were seen in study I, whereas the corrugator EMG responses did not seem to reflect emotions in study V. In this latter study, the corrugator EMG activity decreased significantly after the neutral sentences as compared to the pre-stimulus baseline, whereas the activity did not differ significantly from the pre-stimulus baseline following the negative and positive sentences. Previous research has highlighted how changes in facial muscles reflect not only affective responses, but also a wide variety of non-emotional states and activities. Specifically, the activity of the corrugator EMG facial muscle is associated with concentration and mental effort, in addition to the unpleasant affect (Cacioppo et al., 1985; Van Boxtel & Jessurun, 1993; Smith & Scott, 1997). Some previous work has elucidated how people remember emotionally valenced words better than neutral words probably because of a higher attention allocation to emotionally loaded words (e.g. Doerksen & Shimamura, 2001). Thus, our results suggest that it may be the case that the participants concentrated less on the processing of the neutral sentences than on the processing of the emotionally negative and positive sentences, reflecting more a higher cognitive processing instead of an emotional reaction. However, why did the positive and negative emotional sentences not affect the corrugator EMG responses in study V? One possible explanation is that maybe the changes in emotion were so subtle that there were no differences in EMG activity between the sentence categories. For example, in study I, half of the sentences described the emotional state of a listener. The participants

rated these sentences as significantly more arousing than the sentences that described the emotional state of a computer. Instead, the stimuli in study V consisted only of the sentences that described the emotional state of a computer. Changes in the subjective feelings of emotions can emerge without physiological changes, and vice versa (see Mauss & Robinson, 2009). Specifically, the full-blown prototypical emotions that evoke changes in all elements of emotion (i.e. emotional experience, physiology, and behavior) are quite rare. Emotional states, in which one or more elements are missing, are more common (Russell & Felman Barrett, 1999).

Further, the sentences used in study V were more complex than the sentences used in study I. The study I sentences described the emotional state of the computer or appraised the participant's emotional state straightforwardly (e.g. I am happy; You look angry). The study V sentences were more complex because they contained causative emotional verbs, such as irritation or delight, which express the relation between the cause of an emotion and the person who experiences that emotion (Johnson-Laird & Oatley, 1989). Thus, it might be the case that the study I sentences evoked more primary or basic emotions, such as happiness or anger, and the more complex sentences of study V evoked secondary, social emotions, such as embarrassment or pride. App, McIntosh, Reed, and Hertenstein (2011) noted that the basic emotions are linked primarily to facial behavior, whereas the secondary emotions, which are centrally linked to a person's social interaction, are associated with whole-body behavior.

Finally, facial expressions of emotions are stronger when there is a real or imagined audience present (e.g. Fridlund, 1991). In study I, before the experimental phase, synthesized voices gave comprehensive instructions for the experiment to familiarize the participants with the monotonous voice of speech synthesizers. The instructions were given by the same voices that were also then used in the actual experiment. Instead, in study V, only three short sentences spoken by three different speech synthesizers were presented before the actual experiment to familiarize the participants with listening to synthetic speech. These sentences were spoken by other synthesizers than the one used in the actual experiment. Thus, it is possible that the instructions given by synthesizers in study I evoked a sense of social interaction in the participants, and thus the spoken stimuli used in this experiment evoked emotion-related facial reactions in the participants.

The results of the pupil responses to the affective sentences in study II were in line with much of the previous research in this area. Recent studies have shown that pupils dilate in response to emotionally negative and positive pictures and sounds (Bradley et al., 2008; Partala & Surakka, 2003). In line with these studies, we found that pupils dilated in response

to emotionally negative and positive spoken sentences. Thus, the association between the valence dimension and pupil dilation seems to be curvilinear, as already suggested by Janisse (1974). However, there are some studies in which the valence of visually presented words did not affect pupillary responses (Bayer, Sommer, & Schacht, 2011; Kuchinke, Võ, Hofmann, & Jacobs, 2007). Thus, it might be the case that people process written and spoken material differently. At least there is some evidence that the processing of heard speech takes a little longer than the processing of read words (see Rayner & Clifton, 2009). In addition, it has been suggested that spoken language is a more fundamental means of communication than written language (McGregor, 2009). Altogether, it seems that spoken language can be more powerful at inducing emotion-related pupil responses in people than written words. However, this cannot be confirmed with the present data, because the stimuli only consisted of spoken stimuli. Future research is needed to determine whether the same messages delivered both through the written and spoken language elicit different emotional reactions.

The results of this thesis also show that the processing of spoken language may be different from the processing of pictures. Our heart-rate data in study IV highlighted how the spoken words with emotional content evoked a similar heart-rate response pattern that was found earlier only for longer emotional stimuli. Previous findings about the physiological responses to emotional stimuli have been mainly consistent, regardless of the length of the stimulus. For example, facial EMG changes and skin-conductance responses have been shown to be similar for both brief and longer stimuli (Codispoti, Bradley, & Lang, 2001; Codispoti, Mazzetti, & Bradley, 2009). Conversely, Codispoti et al. (2001) reported that heart-rate responses to briefly presented affective pictures were dramatically different from longer (i.e. six-second) stimuli. They determined that the initial heart-rate deceleration after the presentation of the stimuli was minimal, and, in addition, that picture valence had no significant effect on heart-rate responses. Thus, they suggested that a clear decelerating response to an affective stimulus in the viewing context implies a more sustained sensory stimulation.

However, there are differences in how the different senses process information. The sensory information coming from the environment is initially stored in the sensory memory store (e.g. Whitman, 2011). Different senses process sensory information in different ways, for example, by storing the incoming information differently. For example, auditory sensory information is available for much longer (i.e. seconds) in the sensory storage than is visual information (i.e. only a few hundred milliseconds). The research suggests that the auditory sensory store consists of two phases (Cowan, 1984; 1988). First, the information is stored in an unanalyzed form for several hundred milliseconds. In the second

phase, the content of the information is partly analyzed and it is stored for a few seconds. Thus, the auditory information is still available in the sensory store, while the visual information has progressed to conscious processing in the working memory, and has already faded from the sensory store. The different and slower processing of auditory information may be one explanation for why the hearing of brief lexical stimuli evoked similar heart-rate responses than more continuous visual sensory stimulation. It is also possible that listening binds the attentional resources longer than when watching pictures, because the processing time for verbal material has been shown to be slower than the processing time for pictures (Roelofs, 2008; Schacht & Sommer, 2009b). Taken together, these differences can reflect the fact that the processing of words and pictures is different, and thus, their effects seem to differ from each other.

The results of the last study showed, interestingly, that the content of a message affected how people perceived and rated the speech synthesizers. People rated the voice quality of the synthesizer differently depending on the emotional content of the message. When the content of the sentence was positive, the participants rated the voice as more pleasant and clear than when the content was negative or neutral. Fogg and Nass (1997) determined that people evaluated the interaction with a computer more positively when the computer sincerely or insincerely praised the participant as compared to the situation when the computer gave only some generic feedback. Study V's findings highlighted how the verbal content of the sentences had so strong an effect on people that merely passively listening to the sentences affected the ratings of voice quality, even though the voice remained similar, and even when there was no interaction between the human and the computer.

The results presented in the current thesis suggest that, in general, the emotional content of spoken messages can be used to regulate users' emotions. In addition, positive language, in particular, can be used to enhance positive feelings about the voices. Thus, the results are also interesting from the application point of view, and they can be utilized in interface design. There is strong evidence that positive emotions can affect cognitive processes positively. For example, positive emotions broaden the thought processes, which may be helpful, especially in tasks that require creative thinking. Further, positive emotions facilitate problem-solving behavior and help to keep the motivation level high. Thus, encouraging and motivating messages that aim at positively changing the emotional mind-set of the user can be helpful in situations in which a user benefits from positive emotions. One such suitable environment is within learning applications, in which positive feedback can result in positive learning experiences and effective learning behavior. Bracken and Lombard (2004) noted that in addition to adults, young children also responded socially to computers, and these experiences had consequences in terms of learning

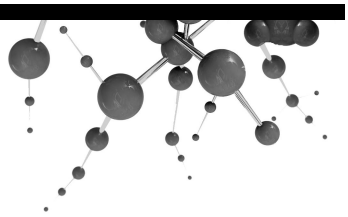
experiences and effectiveness: When children received praising voice messages while completing computerized memory tasks, they rated their own ability higher, and had higher recall and recognition scores than children who received neutral messages. Overall, positive messages can be used to induce positive emotions to optimize the performance level, for example, in learning situations, and they can also motivate people. In addition, because working with computers is cognitively demanding, positive emotional feedback might facilitate recovery from autonomic arousal caused by cognitively demanding tasks, and thus, reduce the user's stress level.

However, there are also situations where a user could benefit from a negative emotional state. As reported in the introduction, Spering et al. (2005) found that positive or negative emotions elicited by giving positive or negative feedback to participants induced different problem-solving strategies. For example, negative emotions led to more detailed and systematic ways of seeking and using information than positive emotions did. While positive emotions broaden the thought processes, negative emotions focus the mind, leading to better concentration (Norman, 2002). Gwizdka and Lopatovska (2009) found that participants who felt less happy before an information-searching task completed the task more thoroughly and felt more in control during the search than the participants who started the task in a happy state. Furthermore, information with negative content may be more effective than positive information when the aim is to catch a user's attention. The research by Bolls, Lang, and Potter (2001) showed that when people listened to radio advertisements their heart rate was slower when the content was negative as compared to positive messages. This finding suggested that negative advertisements received more attention than positive ones did. This is important in respect to our findings on heart-rate changes. The heart rate decelerated the most following exposure to emotionally negative words; thus, even the short, isolated words evoked significant changes in people's heart-rate responses. Consequently, messages coming from a computer do not always have to be positive, as negative messages can function well in some situations.

Study V revealed that the participants rated the quality of the synthesized voice differently depending on the lexical content of the delivered message. Thus, the results also imply that the content of the spoken messages can affect how we evaluate computer interfaces. This knowledge can be utilized when designing different speech interfaces such as automated spoken dialogue systems. For example, positive language can be used to evoke positive feelings when interacting with these kinds of systems.

In this thesis, I have suggested that it is possible to evoke emotion-related reactions and ratings in people through the lexical emotional content of the synthesized spoken messages. Second, I have proposed that increasing the naturalness of the voice can enhance the experienced emotions. The present thesis is limited to only exploring the effects of synthesized voices, but in the future it would be interesting to examine whether there are differences in the reactions to synthesized and human voices. This kind of comparison would give a more thorough understanding of how the level of naturalness affects human emotions. However, in the work described in thesis, the timing and the prosody of the voices were carefully controlled to examine if the lexical content alone was sufficient to affect the emotional reactions, and to compare if the synthesizers or, in other words, the different levels of naturalness would affect the responses. Controlling the prosody of the human voice in a similar manner would be extremely challenging, and for this reason, the stimuli were produced only by speech synthesizers. In the light of the present results, it seems probable that human speech would evoke stronger responses than synthesized speech, but on the other hand, even the monotonous voice produced by speech synthesis seems to be sufficient to evoke emotion-related responses in people.

Furthermore, studying physiological responses requires a controlled environment and a controlled experimental setup, and for this reason, the stimuli in our experiments were presented in a laboratory under controlled conditions. Although the findings may not be directly generalizable to real environments, the results showed quite strong and coherent evidence that synthesized emotional expressions can evoke emotions in people. In our studies, the emotional sentences evoked emotion-related reactions, even though the sentences were presented with a monotonous voice without any interaction context between the human subjects and the computer. Thus, it is probable that the messages delivered in real interaction contexts could evoke even stronger emotions, because then the situation and the messages might have more meaning for people.

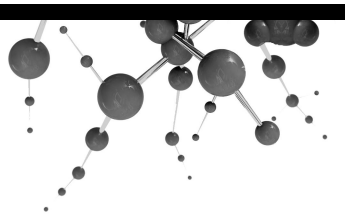


7 Conclusions

In summary, the results presented in this thesis suggest that the synthesized verbal expressions of emotions can evoke emotions in people. Spoken messages evoked emotion-related changes in people's physiology and experiences, even though the messages with lexical emotional content were delivered through monotonous voices. Previous research has suggested that interaction with computers is intrinsically social; that is, people tend to use interaction rules in HCIs that are similar to those that they use in human–human interaction. The present results suggest that computers also evoke emotionality in their users. It seems that the emotional expressions produced by computers could evoke similar emotional responses in humans as the emotional expressions of another human could.

However, the features of the voice are significant. We found evidence that the affective ratings of the stimuli were stronger when the more human-like voices produced the affective stimuli than when the less human-like voice produced the stimuli. Further, only the more human-like voice evoked emotion-related facial muscle and pupil responses when two different synthesizers produced the emotional sentences, as reported in Publications I and II. Thus, in addition to the lexical emotional content, the features of the voice matter when we want to effectively affect the emotions of a computer user.

The results revealed the potentiality of synthetically produced spoken messages in HCIs. The knowledge that speech-based computer systems can induce emotions in people is an important factor to take into account when creating and designing computerized interactive environments and devices.



8 References

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

Anttonen, J. & 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. New York: ACM, 491-499.

Anttonen, J., Surakka, V., & Koivuluoma, M. (2009). Ballistocardiographic responses to dynamic facial displays of emotion while sitting on the EMFi chair. Journal of Media Psychology, 21 (2), 69-84.

App, B., McIntosh, D.N., Reed, C.L., & Hertenstein, M.J. (2011). Nonverbal channel use in communication of emotion: how may depend on why. Emotion, 11 (3), 603-617.

Ashby, F.G., Isen, A.M., & Turken, A.U. (1999). A neuropsychological theory of positive affect and its influence on cognition. Psychological Review, 106 (3), 529-550.

Aula, A. & Surakka, V. (2002). Auditory emotional feedback facilitates human-computer interaction. In Faulkner, X., Finlay, J., & Détienne, F. (Eds.) People and Computers XVI: Memorable Yet Invisible, Proceedings of HCI 2002. London: Springer-Verlag, 337-349.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. Current Directions in Psychological Science, 8 (2), 53-57.

Banse, R. & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70 (3), 614-636.

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

Bayer, M. Sommer, W., & Schacht, A. (2010). Reading emotional words within sentences: the impact of arousal and valence on event-related potentials. International Journal of Psychophysiology, 78 (3), 299-307.

Bayer, M. Sommer, W., & Schacht, A. (2011). Emotional words impact the mind but not the body: evidence from pupillary responses. Psychophysiology, 48 (11), 1553-1561.

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

Bertels, J., Kolinsky, R., & 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.

Bitlips. http://www.bitlips.fi/index.en.html [Accessed 15 August 2012].

Bolls, P.D., Lang, A., & Potter, R.F. (2001). The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. Communication Research, 28 (5), 627-651.

Bracken, C.C. & Lombard, M. (2004). Social presence and children: praise, intrinsic motivation, and learning with computers. Journal of Communication, 54 (1), 22-37.

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. Psychophysiology, 46 (1), 1-11.

Bradley, M.M. & 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., Cuthbert, B.N., & Lang, P.J. (1996). Picture media and emotion: effects of a sustained affective context. Psychophysiology, 33 (6), 662-670.

Bradley, M.M. & 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. (2000a). Measuring emotion: behavior, feeling, and physiology. In Lane, R. & Nadel, L. (Eds.) Cognitive Neuroscience of Emotion. New York: Oxford University Press, 242-276.

Bradley, M.M. & Lang, P.J. (2000b). Affective reactions to acoustic stimuli. Psychophysiology, 37 (2), 204-215.

Bradley, M.M., Miccoli, L., Escrig, M.A., & Lang, P.J. (2008). The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology 45 (4), 602-607.

Brave, S., Nass, C., & 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.

Brosschot, J.F. & Thayer, J.F. (2003). Heart rate response is longer after negative emotions than after positive emotions. International Journal of Psychophysiology, 50 (3), 181-187.

Buchanan, T.W., Etzel, J.A., Adolphs, R., & Tranel, D. (2006). The influence of autonomic arousal and semantic relatedness on memory for emotional words. International Journal of Psychophysiology, 61 (1), 26-33.

Cacioppo, J.T., Berntson, G.G., Klein, D.J., & Poehlman, K.M. (1997). Psychophysiology of emotion across the life span. Annual Review of Gerontology and Geriatrics, 17, 27-74.

Cacioppo, J.T., Berntson, G.C., Larsen, J.T., Poehlmann, K.M., & Ito, T.A. (2000). The psychophysiology of emotion. In Lewis, M. & Haviland, J.M. (Eds.) Handbook of Emotions 2nd Ed. New York: The Guilford Press, 173-191.

Cacioppo, J.T., Petty, R.E., & Morris, K.J. (1985). Semantic, evaluative, and self-referent processing: memory, cognitive effort, and somatovisceral activity. Psychophysiology, 22 (4), 371-384.

Cahn, J.E. (1990). The generation of affect in synthesized speech. Journal of the American Voice IO Society, 8 (1), 1-19.

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

Cannon, W.B. (1929). Bodily changes in pain, hunger, fear and rage. New York: Appleton.

Carver, C.S. & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. Psychological Bulletin, 135 (2), 183-204.

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In Peter, C. & Beale, R. (Eds.) Affect and Emotion in HCI (vol. 3784). Berlin, Heidelberg: Springer, 92-103.

Chappell, D.T. & Hansen, J.H.L. (2002). A comparison of spectral smoothing methods for segment concatenation based speech synthesis. Speech Communication, 36 (3-4), 343-374.

Christie, I.C. & 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.

Coan, J.A., Allen, J.J.B., & Harmon-Jones, E. (2001). Voluntary facial expression and hemispheric asymmetry over the frontal cortex. Psychophysiology, 38 (6), 912-925.

Codispoti, M., Bradley, M.M., & Lang, P.J. (2001). Affective reactions to briefly presented pictures. Psychophysiology, 38 (3), 474-478.

Codispoti, M., Mazzetti, M., & Bradley, M.M. (2009). Unmasking emotion: Exposure duration and engagement. Psychophysiology, 46 (4), 731-738.

Cowan, N. (1984). On short and long auditory stores. Psychological Bulletin, 96 (2), 341-370.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. Psychological Bulletin, 104 (2), 163-191.

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

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J.G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18 (1), 32-80.

Damasio, A.R. (1994). Descartes' error: Emotion, rationality, and the human brain. New York: Putnam (Grosset Books).

Davidson, R.J. (1992). Anterior cerebral asymmetry and the nature of emotion. Brain and Cognition, 20 (1), 125-151.

Dawson, M.E., Schell, A.M., & Filion, D.L. (2000). The electrodermal system. In Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.) Handbook of Psychophysiology 2nd Ed. New York: Cambridge University Press, 200-223.

Darwin, C. (1965). The expression of the emotions in man and animals. Chicago: The University of Chicago Press. (Original work published 1872).

Dehn, D.M. & 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.

De Houwer, J., Baeyens, F., & Eelen, P. (1994). Verbal evaluative conditioning with undetected US presentations. Behaviour Research and Therapy, 32 (6), 629-633.

Doerksen, S. & Shimamura, A.P. (2001). Source memory enhancement for emotional words. Emotion, 1 (1), 5-11.

Dyck, M., Winbeck, M., Leiberg, S., Chen, Y., Gur, R.C., & Mathiak, K. (2008). Recognition profile of emotions in natural and virtual faces. PloS ONE, 3 (11), e3628.

Ekman, P. & Friesen, W. (1978). Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action. Palo Alto, CA: Consulting Psychologists Press.

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

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. Psychological Bulletin, 115 (2), 268-287.

Ekman, P. (1999). Basic Emotions. In Dalgleish, T. & Power, M. (Eds.) Handbook of Cognition and Emotion. Sussex, U.K.: John Wiley & Sons Ltd., 45-60.

Ekman, P., Levenson, R.W., & Friesen, W.V. (1983). Autonomic nervous system activity distinguishes among emotions. Science, 221 (4616), 1208-1210.

Ellsworth, P.C. & Scherer, K.R. (2003). Appraisal processes in emotion. In Davidson, R.J., Goldsmith, H., & Scherer, K.R. (Eds.) Handbook of the

Affective Sciences. New York and Oxford: Oxford University Press, 572-595.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J.T. (2008). When we need a human: a motivational determinants of anthropomorphism. Social Cognition, 26 (2), 143-155.

Epley, N., Waytz, A. & Cacioppo, J.T. (2007). On seeing human: a three-factor theory of anthropomorphism. Psychological Review, 114 (4), 864-886.

Erez, A. & Isen, A.M. (2002). The influence of positive affect on the components of expectancy motivation. Journal of Applied Psychology, 87 (6), 1055-1067.

Estrada, C.A., Isen, A.M., & 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.

Fedurek, P. & Slocombe, K.E. (2011). Primate vocal communication: A useful tool for understanding human speech and language evolution. Human Biology, 83 (2), 153-173.

Feldman Barrett, L. & Russell, J.A. (1999). The structure of current affect: controversies and emerging consensus. Current Directions in Psychological Science, 8 (1), 10-14.

Ferdig, R.E. & Mishra, P. (2004). Emotional responses to computers: experiences in unfairness, anger, and spite. Journal of Educational Multimedia and Hypermedia, 13 (2), 143-161.

Fogg, B.J. & Nass, C. (1997). Silicon sycophants: the effects of computers that flatter. International Journal of Human-Computer Studies, 46 (5), 551-561.

Foroni, F. & Semin G.R., (2009). Language that puts you in touch with your bodily feelings. Psychological Science, 20 (8), 974-980.

Fredrickson, B.L. (2000). Cultivating positive emotions to optimize health and well-being. Prevention & Treatment, 3 (1). Retrieved February 27, 2012, from

http://journals.apa.org/prevention/volume3/pre0030001a.html

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

Fredrickson, B.L. & Levenson, R.W. (1998). Positive emotions speed recovery from the cardiovascular sequelae of negative emotions. Cognition and Emotion, 12 (2), 191-220.

Fridlund, A.J. (1991). Sociality of solitary smiling: potentiation by an implicit audience. Journal of Personality and Social Psychology, 60 (2), 229-240.

Frijda, N.H. (1986). The emotions. Cambridge: Cambridge University Press.

Frijda, N.H. (1989). Relations among emotion, appraisal, and emotional actions readiness. Journal of Personality and Social Psychology, 57 (2), 212-228.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. Computers in Human Behavior, 24 (4), 1494-1509.

Grawitch, M.J., Munz, D.C., Elliott, E.K., & Mathis, A. (2003). Promoting creativity in temporary problem-solving groups: The effects of positive mood and autonomy in problem definition on idea-generating performance. Group Dynamics: Theory, Research, and Practice, 7 (3), 200–213.

Gross, J.J. (2010). The future's so bright, I gotta wear shades. Emotion Review, 2 (3), 212-216.

Guthrie, S.E. (1993). Faces in the Clouds: A New Theory of Religion. New York: Oxford University Press.

Gwizdka, J. & Lopatovska, I. (2009). The role of subjective factors in the information search process. Journal of the American Society for Information Science and Technology, 60 (12), 2452-2464.

Halberstadt, J.B. & Niedenthal, P.M. (2001). Effects of emotion concepts on perceptual memory for emotional expressions. Journal of Personality and Social Psychology, 81 (4), 587-598.

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

Haukkala, A., Konttinen, H., Laatikainen, T., Kawachi, I., & Uutela, A. (2010). Hostility, anger control, and anger expression as predictors of cardiovascular disease. Psychosomatic medicine, 72 (6), 556-562.

Heider, F. & Simmel, M. (1944). An experimental study of apparent behavior. The American Journal of Psychology, 57 (2), 243-259.

Herbert, C., Junghofer, M., & Kissler, J. (2008). Event related potentials to emotional adjectives during reading. Psychophysiology, 45 (3), 487-498.

Herbert, C., & Kissler, J. (2010). Motivational priming and processing interrupt: startle reflex modulation during shallow and deep processing of emotional words. International Journal of Psychophysiology, 76 (2), 64-71.

Herbert, C., Kissler, J., Junghöfer, M., Peyk, P., & Rockstroh, B. (2006). Processing of emotional adjectives: evidence from startle EMG and ERPs. Psychophysiology, 43 (2), 197-206.

Hess, U. & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. International Journal of Psychophysiology, 40 (2), 129-141.

Hietanen, J.K., Surakka, V., & Linnankoski, I. (1998). Facial electromyographic responses to vocal affect expressions. Psychophysiology, 35 (5), 530-536.

Hinojosa, J.A., Carretié, L., Valcárcel, M.A., Méndez-Bértolo, C., & Pozo, M.A. (2009). Electrophysiological differences in the processing of affective information in words and pictures. Cognitive, Affective, & Behavioral Neuroscience, 9 (2), 173-189.

Ho, C.-C., MacDorman, K.F., & Dwi Pramono, Z.A. (2008). Human Emotion and the Uncanny Valley: a GLM, MDS, and Isomap analysis of robot video ratings. Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction. New York: ACM, 169-176

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. Interacting with Computers, 18 (2), 227-245.

Hsee, C.K., Hatfield, E., Carlson, J.G., & Chemtob, C. (1990). The effect of power on susceptibility to emotional contagion. Cognition and Emotion, 4 (4), 327-340.

Hsee, C.K., Hatfield, E., & Chemtob, C. (1992). Assessments of the emotional states of others: Conscious judgments versus emotional contagion. Journal of Social and Clinical Psychology, 11 (2), 119-128.

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. Speech Communication, 40 (1-2), 161-187.

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

Ishii, K., Reyes, J.A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. Psychological Science, 14 (1), 39-46.

Iso-Sipilä, J., Moberg, M., & Viikki, O. (2006). Multilingual speaker-independent voice user interface for mobile devices. In Proceedings of ICASSP 2006, Toulouse, France, 1081-1084.

Izard, C.E. (1977). Human emotions. New York: Plenum Press.

James, W. (1884). What is an emotion? Mind, 9 (34), 188-205.

Janisse, M.P. (1974). Pupil size, affect and exposure frequency. Social Behavior & Personality, 2 (2), 125-146.

Johnson-Laird, P.N. & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. Cognition and Emotion, 3 (2), 81-123.

Kensinger, E.A. & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words. Memory & Cognition, 31 (8), 1169-1180.

Kiecolt-Glaser, J.K., McGuire, L., Robles, T.F., & Glaser, R., (2002). Emotions, morbidity, and mortality: New perspectives from psychoneuroimmunology, 53, 83-107.

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. Journal of Personality and Social Psychology, 70 (1), 47-65.

Kim, Y., Baylor, A.L., & Shen, E. (2007). Pedagogical agents as learning companions: the impact of agent emotion and gender. Journal of Computer Assisted Learning, 23 (3), 220-234.

Kissler, J., Herbert, C., Peyk, P., & Junghofer, M. (2007). Buzzwords: early cortical responses to emotional words during reading. Psychological Science, 18 (6), 475-480.

Kissler, J., Herbert, C., Winkler, I., & Junghofer, M. (2009). Emotion and attention in visual word processing – An ERP study. Biological Psychology, 80 (1), 75-83.

Klatt, D.H. (1987). Review of text-to-speech conversion for English. Journal of Acoustical Society of America, 82 (3), 737-793.

Kleinginna, P. R. & Kleinginna, A. M. (1981). A categorized list of emotion definitions with suggestions for a consensual definition. Motivation and Emotion, 5 (4), 345–379.

Ku, J., Jang, H.J., Kim, K.U., Kim, J.H., Park, S.H, Lee, J.H., Kim, J.J., Kim, I.Y., Kim, S.I. (2005). Experimental results of affective valence and arousal to avatar's facial expressions. CyberPsychology & Behavior, 8 (5), 493-503.

Kuchinke, L., Võ, M.L.-H., Hofmann, M., & Jacobs, A.M. (2007). Pupillary responses during lexical decisions very with words frequency but not emotional valence. International Journal of Psychophysiology, 65 (2), 132-140.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1990). Emotion, attention, and the startle reflex. Psychological review, 97 (3), 377-395.

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

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1997). Motivated attention: Affect, activation, and action. In Lang, P.J., Simons, R.F., & Balaban, M.T. (Eds.) Attention and Orienting – Sensory and Motivational Processes. Mahwah, NJ: Erlbaum, 97-135.

Lang, P.J., Greenwald, M.K., Bradley, M.M., & Hamm, A.O. (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.

Larsen, R.J. & Prizmic-Larsen, Z. (2006). Measuring emotions: implications of a multimethod perspective. In Eid, M. & Diener, E. (Eds.) Handbook of Multimethod Measurement in Psychology. Washington DC, USA: American Psychological Association, 337-351

Law, E.L.-C., Roto, V., Hassenzahl, M., Vermeeren, A.P., & Kort, J. (2009). Understanding, scoping and defining user eXperience: A survey approach. In Proceedings of CHI'09. New York: ACM, 417-435.

Levenson, R.W. (1992). Autonomic nervous system differences among emotions. Psychological Science, 3 (1), 23-27.

Levenson, R.W., Ekman, P., & Friesen, W.V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. Psychophysiology, 27 (4), 363-384.

Lisetti, C.L. & Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. Journal on Applied Signal Processing, 11, 1672-1687.

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

Mayer, J.D. & Salovey, P. (1997). What is emotional intelligence? In Salovey, P. & Sluyter, D.J. (Eds.) Emotional Development and Emotional Intelligence. New York: BasicBooks, 3-31.

Maynard Smith, J. & Száthmáry, E. (1997). The major transitions in evolution. New York: Oxford University Press.

McGregor, W. (2009). Linguistics: An Introduction. New York: Continuum International Publishing Group.

Mikropuhe©. http://www.mikropuhe.com/ [Accessed: 15 August 2012].

Miller, G.A. (1981). Language and speech. San Francisco: W.H. Freeman and Company.

Mishra, P. (2006). Affective feedback from computers and its effect on perceived ability and affect: A test of the computers as social actor hypothesis. Journal of Educational Multimedia and Hypermedia, 15 (1), 107-131.

Misselhorn, C. (2009). Empathy with inanimate objects and the uncanny valley. Minds & Machines, 19 (3), 345-359.

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

Murray, I.R. & 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. & Brave, S. (2005). Wired for speech: How voice activates and advances the human-computer relationship. Cambridge: The MIT Press.

Nass, C., Fogg, B.J., & Moon, Y. (1996). Can computers be teammates? International Journal of Human-Computer Studies, 45 (6), 669-678.

Nass, C. & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56 (1), 81-103.

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. Journal of Applied Social Psychology, 29 (5), 1093-1110.

Nass, C., Moon, Y., & Green N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. Journal of Applied Social Psychology, 27 (10), 864-876.

Nass, C. & Lee, K.W. (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.

Ng, S.K. & Bradac, J.J. (1993). Power in language: verbal communication and social influence. Newbury Park, CA: Sage Publications.

Norman, D.A. (2002). Emotion & design: attractive things work better. Interactions, 9 (4), 36-42.

Nowak, K.L. (2004). The influence of anthropomorphism and agency on social judgment in virtual environments. Journal of Computer-Mediated Communication, 9 (2) [online]. Available from: http://www3.interscience.wiley.com/cgi-bin/fulltext/120837918/HTMLSTART [Accessed: 23 March 2012].

Nowak, K.L. & 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. & 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.

Nygaard, L.C. & Queen, J.S. (2008). Communicating emotion: Linking affective prosody and word meaning. Journal of Experimental Psychology: Human Perception and Performance, 34 (4), 1017-1030.

Ochs, M., Niewiadomski, R., Pelachaud, C. & Sadek, D. (2005). Intelligent expressions of emotions. In J. Tao, T. Tan, & R. Picard (Eds.) Affective Computing and Intelligent Interaction (vol. 3784). Berlin, Heidelberg: Springer, 707-714.

Ochsner, K.N. & Feldman Barrett, L. (2001). A multiprocess perspective on the neuroscience of emotion. In Mayne, T. & Bonnano, G. (Eds.)

Emotion: Current Issues and Future Directions. New York: Guilford Press, 38-81.

Ortony, A. & Turner, T.J. (1990). What's basic about basic emotions? Psychological Review, 97 (3), 315-331.

Palomba, D., Angrilli, A., & Mini, A. (1997). Visual evoked potentials, heart rate responses and memory to emotional pictorial stimuli. International Journal of Psychophysiology, 27 (1), 55-67.

Partala, T. & Surakka, V. (2003). Pupil size variation as an indication of affective processing. International Journal of Human-Computer Studies, 59 (1-2), 185-198.

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

Partala, T., Surakka, V. & Vanhala, T. (2006). Real-time estimation of emotional experiences from facial expressions. Interacting with Computers, 18 (2), 208-226.

Phan, K.L, Wager, T., Taylor, S., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage 16 (2), 331-348.

Phan, K.L, Wager, T., Taylor, S., & Liberzon, I. (2004). Functional neuroimaging studies of human emotions. CNS Spectrums, 9 (4), 258-266.

Picard, R. W. (1997) Affective computing. Cambridge: The MIT Press.

Picard, R.W. (2003). Affective computing: challenges. International Journal of Human-Computer Studies, 59 (1-2), 55-64.

Picard, R., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions of Pattern Analysis and Machine Intelligence, 23 (10), 1175-1191.

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 59 (1-2), 157-183.

Pinker, S. (1995). The language instinct: the new science of language and mind. London: Penguin Books.

Plutchik, R. (1980). Emotion. A psychoevolutionary synthesis. New York: Harper & Row, Publishers.

Posner, J., Russell, J.A., & Peterson, B.S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17 (3), 715-734.

Power, G., Wills, G., & Hall, W. (2002). User perception of anthropomorphic characters with varying levels of interaction. In Faulkner, X., Finlay, J., & Détienne, F. (Eds.) People and Computers XVI: Memorable Yet Invisible, Proceedings of HCI 2002. London: Springer-Verlag, 37-52.

Prendinger, H., Becker, C., & Ishizuka, M. (2006). A study in users' physiological response to an empathic interface agent. International Journal of Humanoid Robotics, 3 (3), 371-391.

Prendinger, H., Mori, J., & 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.

Pressman, S.D. & Cohen, S. (2005). Does positive affect influence health? Psychological Bulletin, 131 (6), 925-971.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A.R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. International Journal of Psychophysiology, 61 (1), 5-18.

Rayner, K. & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. Biological Psychology, 80 (1), 4-9.

Reeves, B. & Nass, C. (1996). The media equation: how people treat computers, television, and new media like real people and places. California: CSLI Publications.

Robinson, M.D. & Clore, G.L. (2002a). Belief and feeling: evidence for an accessibility model of emotional self-report. Psychological Bulletin, 128 (6), 934-960.

Robinson, M.D. & Clore, G.L. (2002b). Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. Journal of Personality and Social Psychology, 83 (1), 198-215.

Roelofs, A. (2008). Dynamic of the attentional control of word retrieval: analyses of response time distributions. Journal of Experimental Psychology: General, 137 (2), 303-323.

Rosenkranz, M.A., Jackson, D.C., Dalton, K.M., Dolski, I., Ryff, C.D., Singer, B.H., Muller, D.,

Kalin, N.H., & Davidson, R.J. (2003). Affective style and in vivo immune response: Neurobehavioral mechanism. Proceedings of the National Academy of Sciences, 100 (19), 11148-11152.

Russell. J.A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39 (6), 1161-1178.

Russell, J.A. & Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of Personality and Social Psychology, 76 (5), 805-819.

Saarni, T. (2010). Segmental durations of speech. Doctoral dissertation, University of Turku, Finland. Available from http://www.doria.fi/handle/10024/52552 [Accessed: 15 August 2012].

Schacht, A. & Sommer, W. (2009a). Time course and task dependence of emotion effects in word processing. Cognitive, Affective, & Behavioral Neuroscience, 9 (1), 28-43.

Schacht, A. & Sommer, W. (2009b). Emotions in word and face processing: Early and late cortical responses. Brain & Cognition, 69 (3), 538-550.

Schachter, S. & Singer, J.E. (1962). Cognitive, social, and physiological determinants of emotional state. Psychological review, 69 (5), 379-399.

Scharf, B. (1998). Auditory attention: The psychoacoustical approach. In Pashler, H. (Ed.) Attention. Hove: Psychology Press, 75-117.

Schaumburg, H. (2001). Computers as tools or social actors? – The users' perspective on anthropomorphic agents. International Journal of Cooperative Information Systems, 10 (1 & 2), 217-234.

Scherer, K.R. (1986). Vocal affect expression: A review and a model for future research. Psychological Bulletin, 99 (2), 143-165.

Scherer, K.R. (1989). Vocal measurement of emotion. In Plutchik, R. & Kellerman, H. (Eds.) Emotion: Theory, Research, and Experience, Vol. 4. New York: Academic Press, 233-259.

Scherer, K.R., Banse, R., Wallbot, H.G., & Goldbeck (1991). Vocal cues in emotion encoding and decoding. Motivation and Emotion, 15 (2), 123-148.

Schirmer, A., Zysset, S., Kotz, S.A., & von Cramon, Y. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. NeuroImage, 21 (3), 1114-1123.

Schlosberg, H. (1954). Three dimensions of emotion. The Psychological Review, 61 (2), 81-88.

Schröder, M. (2009). Approaches to emotional expressivity in synthetic speech. In Izdebski, K. (Ed.) Emotions in the Human Voice. Volume III: Culture and Perception. San Diego: Plural Publishing Inc., 307-321.

Schluroff, M. (1982). Pupil responses to grammatical complexity of sentences. Brain & Language, 17 (1), 133-145.

Scott, G.G., O'Donnell, P.J., Leuthold, H., & Sereno, S.C. (2009). Early emotion word processing: evidence from event-related potentials. Biological Psychology, 80 (1), 95-104.

Seyama, J. & Nagayama, R.S. (2007). The uncanny valley: effect of realism on the impression of artificial human faces. Presence, 16 (4), 337-351.

Silvert, L., Delplanque, S., Bouwalerh, H., Verpoort, C., & Sequeira, H. (2004). Autonomic responding to aversive words without conscious valence discrimination. International Journal of Psychophysiology, 53 (2), 135-145.

Smith, G.A. & Scott, H.S. (1997). A componential approach to the meaning of facial expressions. In Russell, J.A. & Fernandez-Dols J.M. (Eds.) The psychology of facial expressions. New York: Cambridge University Press, 229-254.

Spering, M., Wagener, D., & Funke, J. (2005). The role of emotions in complex problem-solving. Cognition and Emotion, 19 (8), 1252-1261.

Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., & Waters, K. (1996). When the interface is a face. Human-Computer Interaction, 11 (2), 97-124.

Suopuhe. http://www.ling.helsinki.fi/suopuhe/english.shtml [Accessed: 15 August 2012].

Surakka, V. & Hietanen, J.K. (1998). Facial and emotional reactions to Duchenne and non-Duchenne smiles. International Journal of Psychophysiology, 29 (1), 23-33.

Surakka, V. & Vanhala, T. (2011). Emotions in human-computer interaction. In Kappas, A. & Krämer, N. (Eds.) Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language and Technology. Cambridge: Cambridge University Press, 213-236.

Tinwell, A. & Grimshaw, M. (2009). Bridging the Uncanny: an impossible traverse? In Sotamaa, O., Lugmayr, A., Franssila, H., Näränen, P., Vanhala, J. (Eds.) 13th International MindTrek

Conference: Everyday Life in the Ubiquitous Era. New York: ACM, 66-73.

Tinwell. A., Grimshaw, M., & Williams, A. (2011). The Uncanny Wall. International Journal of Arts and Technology, 4 (3), 326-341.

Van Boxtel. A. & Jessurun, M. (1993). Amplitude and bilateral coherency of facial and jaw-elevator EMG activity as an index of effort during a two-choice serial reaction task. Psychophysiology, 30 (6), 589-604.

Van Den Hout, M.A., De Jong, P., & Kindt, M. (2000). Masked fear words produce increased SCRs: An anomaly for Öhman's theory of pre-attentive processing in anxiety. Psychophysiology, 37 (3), 283-288.

Van Gerven, P.W.M., Paas, F., Van Merriënboer, J.J.G., & Schmidt, H.G. (2004). Memory load and the cognitive pupillary response in aging. Psychophysiology, 41 (2), 167-174.

Van Lancker Sidtis, D. (2008). The Relation of Human Language to Human Emotion. In Stemmer, B. & Whitaker, H.A. (Eds.) Handbook of the neuroscience of language. Netherlands, Amsterdam: Elsevier, 199-207.

Vanhala, T., Surakka, V., Siirtola, H., Räihä, K.-J., Morel, B, & Ach, L. (2010). Virtual proximity and facial expressions of computer agents regulate human emotions and attention. Computer Animation and Virtual Worlds, 21 (3-4), 215-224.

Verney, S.P., Granholm, E., & Dionisio, D. (2001). Pupillary responses and processing resources on the visual backward masking task. Psychophysiology, 38 (1), 76-83.

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: 15 June 2012].

Waldstein, S.R., Kop, W.J., Schmidt, L.A., Haufler, A.J., Krantz, D.S., & Fox, N.A. (2000). Frontal electrocortical and cardiovascular reactivity during happiness and anger. Biological Psychology, 55 (1), 3-23.

Wexler, B.E., Warrenburg, S., Schwarz, G.E., & Janer, L.D. (1992). EEG and EMG responses to emotion-evoking stimuli processed without conscious awareness. Neuropsychologia, 30 (12), 1065-1079.

Weyers, P., Mühlberg, A., Hefele, C., & Pauli, P. (2006). Electromyographic responses to static and dynamic avatar emotional facial expressions. Psychophysiology, 43 (5), 450-453.

Whalen, P.J., Rauch, S.L., Etcoff, N.L., McInerney, S.C., Lee, M.B., & Jenike, M.A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. The Journal of Neuroscience, 18 (1), 411-418.

Whitman, D. (2011). Cognition. United States of America: John Wiley & Sons Inc.

Wickens, C.D. (2002). Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, 3 (2), 159-177.

Wundt, W. (1896). Outlines of psychology. Leipzig: Wilhelm Engelmann.

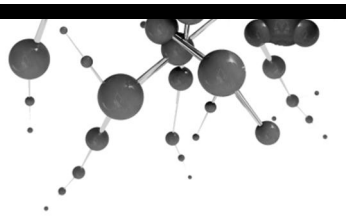
Wurm, L.H. & Vakoch, D.A. (1996). Dimensions of speech perception: Semantic associations in the affective lexicon. Cognition and Emotion, 10 (4), 409-423.

Yee, N., Bailenson, J.N., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interface. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, 1-10.

Yiend, J. (2010). The effects of emotion on attention: A review of attentional processing of emotional information. Cognition and Emotion, 24 (1), 3-47.

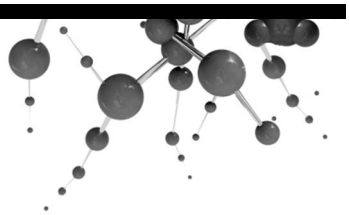
Yik, M.A.M., Russell, J.A., & Feldman Barrett, L. (1999). Structure of self-reported current affect: integration and beyond. Journal of Personality and Social Psychology, 77 (3), 600-619.

Zajonc, R.B. (1980). Feeling and thinking: Preferences need no inferences. American Psychologist, 35, 151-175.



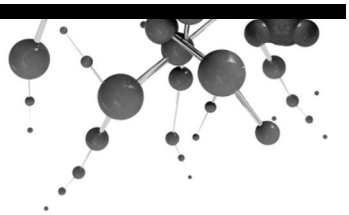
Publication I

Ilves, M. and Surakka, V. (2004). Subjective and physiological responses to emotional content of synthesized speech. *Proceedings of the International Conference on Computer Animation and Social Agents, CASA2004 (Geneva, Switzerland)*, July 2004, Computer Graphics Society, pages 19-26.



Publication II

Ilves, M. and Surakka, V. (2009). Emotions, anthropomorphism of speech synthesis, and psychophysiology. In Izdebski, K. (Ed.) *Emotions in the human voice. Volume III: Culture and perception*. San Diego, USA: Plural Publishing, Inc., pages 137-152.

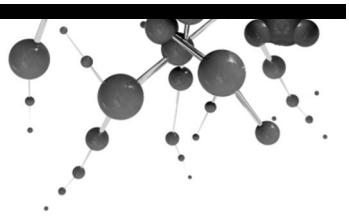


Publication III

Ilves, M. and Surakka, V. (2013). Subjective responses to synthesised speech with lexical emotional content: the effect of the naturalness of the synthetic voice. *Behaviour & Information Technology*, 32 (2), 117-131.

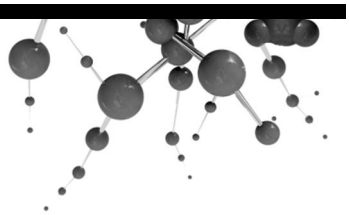
Original article available online at:
http://dx.doi.org/10.1080/0144929X.2012.702285

Publication IV



Ilves, M. and Surakka, V. (2012). Heart rate responses to synthesized affective spoken words. *Advances in Human-Computer Interaction*, 2012, article ID 158487.

Original article available online at:
http://dx.doi.org/10.1155/2012/158487



Publication V

Ilves, M., Surakka, V., and Vanhala, T. (2011). The effects of emotionally worded synthesized speech on the ratings of emotions and voice quality. In *Affective Computing and Intelligent Interaction (ACII 2011), Part I, Lecture Notes in Computer Science*, 6974. Springer-Verlag, pages 588-598.

1. Timo Partala: Affective Information in Human-Computer Interaction
2. Mika Käki: Enhancing Web Search Result Access with Automatic Categorization
3. Anne Aula: Studying User Strategies and Characteristics for Developing Web Search Interfaces
4. Aulikki Hyrskykari: Eyes in Attentive Interfaces: Experiences from Creating iDict, a Gaze-Aware Reading Aid
5. Johanna Höysniemi: Design and Evaluation of Physically Interactive Games
6. Jaakko Hakulinen: Software Tutoring in Speech User Interfaces
7. Harri Siirtola: Interactive Visualization of Multidimensional Data
8. Erno Mäkinen: Face Analysis Techniques for Human-Computer Interaction
9. Oleg Špakov: iComponent – Device-Independent Platform for Analyzing Eye Movement Data and Developing Eye-Based Applications
10. Yulia Gizatdinova: Automatic Detection of Face and Facial Features from Images of Neutral and Expressive Faces
11. Päivi Majaranta: Text Entry by Eye Gaze
12. Ying Liu: Chinese Text Entry with Mobile Devices
13. Toni Vanhala: Towards Computer-Assisted Regulation of Emotions
14. Tomi Heimonen: Design and Evaluation of User Interfaces for Mobile Web Search
15. Mirja Ilves: Human Responses to Machine-Generated Speech with Emotional Content