The Influence of the Auditory Environment on the Emotional Perception of Speech

Maarten Brouwers, 2008

0585320

Master Thesis

Graduation committee

prof.dr. Armin Kohlrausch (TU/e – TIW/HTI / Philips Research)

dr. Dik Hermes (TU/e - TIW/HTI)

dr.ir. Harm Belt (Philips Research)

Human Technology Interaction

Technologie Management

Technische Universiteit Eindhoven

May, 2008

Keywords: EMOTIONS, CONTEXT, COMMUNICATION, AUDIO, PERCEPTION, SPEECH, ENVIRONMENT, AMBIENT

Preface

About seven months ago I started the work, that I describe here in this thesis, at Philips Research. Just before the start of the summer holiday break, Armin Kohlrausch gave me a short description of an open intern position at Philips Research. Some questions and directions were formulated related to the added value of spatial sound in mediated friends and family communication, but after seeing the demo of the system, I was fascinated by the noise suppressing technique that was demonstrated. After some thinking, and reviewing literature, a new question was formulated in which the effect of environmental audio was the subject of study. I am very happy with the freedom I was given by Armin Kohlrausch and Harm Belt at Philips Research when developing this thesis. And although it was not what they were asking for in the first place, it did not seem to affect their support.

Additionally, I want to thank Dik Hermes, who has been a great teacher, but was also a good supporter along the process. Thank you, not only for what you have done during this final thesis work, including the creation of the close-copy contours, but also the earlier mentoring and teaching was very valuable to me.

Many thanks also to Emiel Krahmer and Marc Swerts from the University of Tilburg, who were apparently former colleagues of Armin Kohlrausch and Dik Hermes, and by accident also interested in quite similar questions related to multi-modal perception, emotions in speech, etc. Thanks to Emiel and Marc, much more attention was paid to the generation of the stimuli. Thank you for introducing to me the Velten method, providing me with the film stimuli, helping in finding participants, and the feedback during the process.

But besides 'work' there is also a private life. My love goes out to my girlfriend Mariëlle, who supported besides me, her mother, who was found to be seriously ill halfway this graduation project. I am still feeling sorry for spending so much time on writing and working on my thesis, instead of supporting you, in this difficult period. I want to wish her and her family all best. Luckily there was also Leicester, who is always there to cheer up everyone, taking us out for a walk in the country side, or simply told us: “play ball [translated, ed.]” .

Many thanks, of course, go out to my parents, who have been supporting me in all my

choices I made. Without their support, I wouldn't have been where I am today. Thank you very much!

And then there are the participants, friends, other family, colleagues at Philips Research, interns at Philips in particular... thank you all (and my apologies for not mentioning you all by name).

Maarten Brouwers
- May, 2008

Summary

In this thesis the influence of the auditory environment on the emotional perception of speech in mediated communication is addressed. The motivation of this study is the development of techniques that enable suppression of environmental sound, with the goal to increase the signal-to-noise ratio in communication devices, and thus improving speech intelligibility.

Based on the commotion model of Scherer and Zentner (2001), two routes are discerned in which the auditory environment is thought to affect the emotional percept of a sender. A direct route, changing the interpretation based on the auditory environment heard, and an indirect route, changing the way in which a sender speaks, and thereby the prosodic features that can be perceived and understood as symptoms of emotion. Two hypotheses were formulated:

1. An auditory environment with a distinct emotional quality influences the perceived emotional quality of emotional speech;
2. Speech recorded in noisy environments is perceived less neutral when listened to outside the noisy context; i.e. more negative and more aroused.

In Experiment 1, stimuli were generated with two emotion induction methods, namely the Velten-mood induction procedure and a film-based induction procedure. In Experiment 2, a selection of the stimuli was made. In experiments 3 and 4 hypotheses 1 and 2 were tested.

Evidence was only found for the first hypothesis. To test hypothesis 1, emotional (positive, negative, neutral) utterances were played back in different (emotional) environments in Experiment 3. The environments significantly influenced the emotional perception of the sentences. In case of a positively perceived environment, utterances were in general perceived as more positive, too, whereas the negative environments lead to a more negative percept. On the emotional dimension of valence, the environmental sounds seemed to shift the perceived valence of the utterance towards the valence of the environment. On the the emotional dimension of arousal, however, the perceived arousal of the utterance seemed

to decrease when arousal of the environment increased, suggesting a shift away from the environment's emotional content. While confirmed in Experiment 4 for the neutral speech used in that experiment, the shift was not significant for the aroused Lombard speech, speech recorded in a noisy environment. The second hypothesis was, therefore, rejected.

Based on these results, it is not expected that suppressing the environmental sound in communication will lead to serious misunderstandings due to misinterpretation of a sender's emotion. Environmental noise does not seem to affect the perceived arousal of already aroused speech, nor the valence of the sender's emotion. Additionally, the attracting effect of the positive and negative contexts on the valence axis, the significant effect found for the environment resulting in a relatively neutral percept of someone's emotion when presented in isolation, is not expected to lead to major misunderstandings either.

Samenvatting

In dit afstudeerverslag wordt de invloed van geluidsomgevingen op de emotionele perceptie van spraak in gemedieerde communicatie bestudeerd. De ontwikkeling van technieken die onderdrukking van omgevingsgeluid mogelijk maken, met het doel om de signaal-ruisverhouding in communicatie systemen, en daarmee de verstaanbaarheid, te verbeteren, vormden de aanleiding tot dit onderzoek.

In het 'commotie' (of meeleven) model van Scherer en Zentner (2001) worden twee routes onderscheiden waarin omgevingsgeluid de emotionele perceptie van de spreker beïnvloedt. Een directe route waarbij de interpretatie van de ontvanger wordt beïnvloedt door het omgevingsgeluid, en een indirecte route waarbij het omgevingsgeluid de manier van spreken beïnvloedt van de spreker, en daarmee de prosodische eigenschappen van de spreker, welke eventueel ook als emotioneel kunnen worden waargenomen. Twee hypotheses werden opgesteld:

1. Een auditieve omgeving met een duidelijke emotionele inhoud beïnvloedt de waarneming van de emotionele inhoud van emotionele spraak
2. Spraak opgenomen in een rumoerige omgeving, wordt minder neutraal ervaren wanneer het zonder het rumoer van de omgeving wordt gehoord: spraak wordt dan waargenomen als negatiever, en vooral actiever.

Om de eerste hypothese te testen werden in het eerste experiment stimuli gegenereerd met behulp van twee emotie-inductiemethoden: de methode van Velten en inductie met behulp van filmfragmenten. In het tweede experiment werd een selectie gemaakt uit deze fragmenten. In het derde en vierde experiment werden vervolgens respectievelijk de eerste en de tweede hypothese getoetst.

Alleen de eerste hypothese werd bevestigd. Om deze eerste hypothese te testen werden emotioneel uitgesproken zinnen (positief, negatief, neutraal) afgespeeld in verschillende (positief, negatief, stilte, lawaaiig, extreem lawaaiig) omgevingen. In het experiment was de invloed van de omgeving significant. Binnen een positieve omgeving

werden zinnen over het algemeen als positiever ervaren, terwijl binnen de negatieve omgeving de zinnen als negatiever werden ervaren. Hoewel er een aantrekking uitging van de emotionele omgevingen op de valentie component (positief-negatief) van een emotie, leken de verschillende lawaaiige omgevingen, die als actief en actiever werden beoordeeld, een afstotend effect te hebben op de emotionele waarneming waar het de activatie component (actief-passief) betreft. Hoewel in experiment 3 dit effect werd bevestigd, bleek de in een rumoerige omgeving opgenomen spraak (zgn. Lombard spraak), zoals getest in experiment 4, verrassend genoeg stabiel voor het al dan niet presenteren van de rumoerige context (hetgeen ertoe leidde dat de tweede hypothese werd verworpen).

Op basis van deze resultaten wordt niet verwacht dat het onderdrukken van omgevingsgeluid geen misverstanden zal veroorzaken als gevolg van een foutieve interpretatie van de emotie van de spreker. Omgevingslawaai lijkt niet de waargenomen activatie van de spreker te beïnvloeden (ook in rumoerige context wordt deze al als geactiveerd waargenomen). Tevens is het lastig voor te stellen dat het ietwat neutraliserende effect door het weg laten van positief of negatief omgevingsgeluid leidt tot misverstanden.

Table of Contents

1. Introduction

Communication between persons often involves communication of emotion. Much of this emotion is not clearly stated in words, but transferred in the vocal properties of speech. The way words are pronounced may tell something about how a speaker feels. It may communicate, for example, that a person is rather bored with the conversation, or, quite to the contrary, is seriously interested and enthusiastic. There is much evidence supporting the idea that people are capable of recognizing someone's emotion by only hearing an isolated utterance (see for an overview Scherer, 2003), quite comparable with our capability of 'reading one's emotion' from a facial expressions. Many of these experiments, however, were conducted in controlled situations where only isolated utterances were presented to the listener. In everyday life, most communication takes place in environments which are full of other sounds. These additional sounds may make communication much more difficult. Consider for example communicating from a party: people have to talk louder than normal to make themselves understood. On the other hand, however, these sounds can provide information about the situation or environment in which the communication takes place; the environment thereby creating a context for the conversation, which may help in understanding what is said. It may not surprise you when you hear people talking enthusiastically at a party, but imagine that you would hear the same people talking in the same enthusiastic way at a funeral, with crying people in the background or simply in silence. And how do people interpret someone talking with a plain voice in a party context? Central to these issues is the influence of the environment on the emotional perception on speech. The central question in this thesis therefore is: “How do different auditory environments influence how persons are perceived emotionally?”

1.1 Communication of Emotion

Much of the literature on emotion is about how we perceive and recognize emotions. It addresses how we recognize emotional expressions in, e.g., faces and speech. How we perceive emotional experiences of others, however, is largely neglected (considering the relatively small amount of literature available on this topic).

To explain the ideas in this thesis in more detail, a model created by Scherer and Zentner (2001; based on Scherer, 1998) will be used. This model describes the process of perceiving (and understanding) a person talking with some affective quality in a certain

BOX 1: EMOTION

Emotion is often considered as some internal state, a result of an appraisal process. Such process constitutes earlier occurred situations or event(s), which are appraised by an observer (alternatively: 'evaluated' or 'affectively judged' (see Russell, 2003, p. 149)), and influence, to a variable extent, the emotional response. This emotional response can be a subjective experience, resulting in physiological changes internally, but may also be reflected in externally perceivable symptoms (Rosenberg, 1998; Scherer and Zentner, 2001; Smith, Nolen-Hoeksema, Fredrickson, Loftus, 2003).

The quality of an object that is capable of triggering an affective response is called the *affective quality* of an object (Russell, 2003). Russell's conceptualisation of affective quality resembles to a large extent how humans are thought to perceive other qualities, e.g. affordances like sitability and eatability, etc. (it 'just is')

The emotional experience that follows from perceiving something with an affective quality can be broken down into several aspects. The most important being referred to as *core affect*. Core affect is defined as the "contentful state of pleasure or displeasure" (Barrett, Mesquita, Ochsner & Gross, 2007, p. 377) and is considered as something universal, and present in all humans from birth. Core affect alone, however, is not sufficient to represent and describe an emotional experience entirely. Aside from core affect, Barrett et al. mention arousal content, relational content and the situational content as important means to discern between several emotional experiences. And even these appraisal dimensions may not always be sufficient to describe every type of emotional experience.

Arousal relates to feelings like excitement and activation versus a more sleepy, quiet state (Barrett et al., 2007). Some theorists, among whom Russell (2003), believe arousal content should be considered as part of the core emotional experience, however, there is no solid evidence for a one-to-one mapping between physical activation and felt arousal (Barrett et al., 2007).

Relational content refers to the incorporation of others, and the relationship between oneself and these others (e.g. status, respect) in a mental representation of an emotion. (Barrett et al., 2007).

Situational content, finally, refers to the meaningful content of a situation to the appraiser (Barrett et al., 2007), thus whether it is a) novel or unexpected or not, b) obstructive to one's goals or not, c) compatible or not with norms and values, d) something one is responsible for, or not. According to Barrett et al. about half the variance that differentiates categories of emotion experiences can be explained with core affect and situational content alone.

Situation and relation are also an important aspects in Lazarus' core relational themes or appraisal patterns (Lazarus 1991; via Smith, Nolen-Hoeksema, Fredrickson, Loftus, 2003). For example, disgust can be explained as 'taking in or being too close to an indigestible object or idea' (p. 394).

It is clear that there is no definitive agreement on the exact definition of emotion, nor how to describe an emotional experience. Because of that, also measuring emotions is probably never perfect. This issue, however, will be addressed later in this thesis in more detail.

setting. Originally this model was developed for modelling the process of commotion in mediated unidirectional events, like movies or news-items (Scherer, 1998). But it has also been used in modelling how a music performance is understood emotionally (Scherer and Zentner, 2001). Though rather simple, this model allows for describing the basic anticipated effects of context on emotion in two way communication settings.

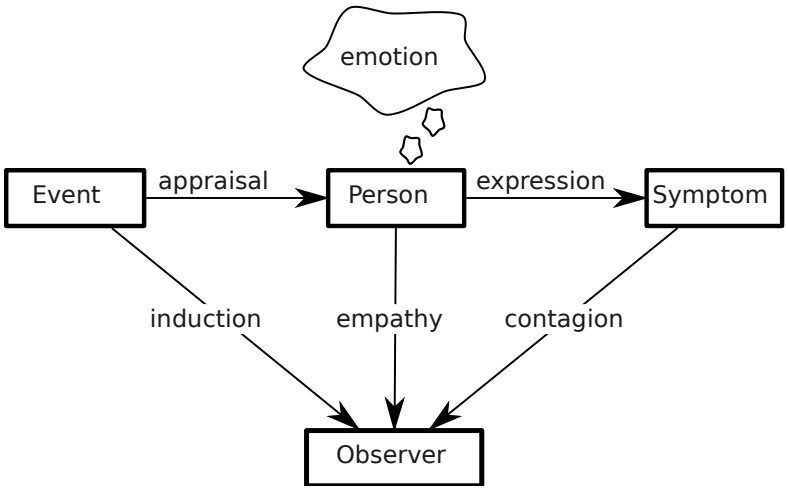

Figure 1 . Commotion (adapted from Scherer, 1998; Scherer & Zentner, 2001). The cloud represents a to the observer hidden construct of the emotion experienced by the person sending which cannot be perceived.

In the model of commotion (Figure 1), commotion is thought to be caused by the processes of induction, empathy and contagion. Induction refers to how people use knowledge about the event in interpreting the entire situation, a process which may involve a process of reasoning how they would feel in that situation. Empathy refers to the relation with and/or the attitude towards the person sending. Empathy is expected to mediate the observer's feeling towards this experience. Liking or disliking the sender, for example, may lead to different conclusions. Contagion, lastly, refers to the process in which externally observable symtoms, e.g. the prosodic features when speech is concerned, influence the observer's commotion.

Two routes can be discerned in the model in which the environment influences commotion of the receiver, see Figure 1. The first route that will be discussed is the direct influence of the environment on the receiver (the induction route). Followed by a discussion of the influence of the environment on the speaker talking (that of appraisal), which influences the commotion via the change in expression. The empathy route will not be considered in this thesis, as it is considered relatively stable.

1.2 Environment influencing the perceived emotion

The first influence of the environment discerned is the direct influence of the context (Scherer, 1998; Scherer & Zentner, 2001). The environment directly influences how the emotions of a person are being perceived.

The fact that changes to the perceived situation can influence the emotional percept of relatively neutral utterances has been demonstrated by Cauldwell (2000). In two experiments large changes in the judgement of the perceived emotion of anger were found. Without any context a single utterance “What do you mean?” was interpreted by 66% as angry, whereas after being told about the relatively peaceful situational context, a father talking to his son, and temporal context, a transcript of the entire conversation, a significant shift in the perceived angriness was found. After presentation of the context, only 10% of the people rated the utterance as angry. Cauldwell suggested that this outcome might have to do with what we experience as being a normal voice: everything that deviates from this is considered not neutral. If this explained the results, a repetition of the experiment, allowing for a habituation period should not lead to similar results. In a follow up experiment he therefore reversed the order: first the entire recording was presented, from which the context could be

BOX 2: CONTEXT DEFINED

The definition of context used in this document can be loosely interpreted as everything but the central subject, and is therefore used interchangeably with 'event', 'situation' or 'environment', which is from the observer's position all context; i.e. gives more information about the subject. Note that from the sender's perspective the event with the emotional quality capable of changing his or hers emotional state is probably not perceived as context.

The term context is not used consistently by different researchers. Douglas-Cowie, Campbell, Cowie and Roach (2003), for example, view context quite differently, and discern: semantic context, structural context, inter-modal context and temporal context, of which only the last type, temporal context, is considered context as it was defined in the previous paragraph.

Semantic, structural and intermodal content all refer to signals or symptoms originating directly from the sender. With semantic context, Douglas-Cowie et al. refer to the emotion contained in the language itself, structural context to syntactic structures and other variations in the speech prosody, and inter-modal context, lastly to the integration of inputs from different modalities (e.g. Massaro & Egan, 1996 (voice & face); de Gelder & Vroomen, 2000 (voice & face); Van den Stock, Righart & de Gelder, 2007 (body & face)).

Temporal context refers to how time may influence the interpretation; how what is said before or after influences the perceived emotion. When somebody recalls something funny just before the emotional expression, the emotional expression might relate to this memory.

More related to the conception context used in this document are what Barrett, Mesquita, Ochsner and Gross refer to as relational context, e.g. who is involved in the interaction (p. 377), and situational context (e.g. Barrett et al., 2007; Scherer & Zentner; 2001)

deduced, which was followed by another trial in which only the isolated words were played. Even in this order, a similar shift was recorded, but now the other way around: first, with context, only 13 % rated the utterance as utterance as angry, whereas without context 72% rated it as angry. This result was quite like the result obtained in the previous experiment, and therefore habituation to the voice may not be the best explanation. The results support the possibility that perceived emotions are strongly influenced by a change in context.

Figure 2 . A anchor face (left) next to a target face (right). In Russel and Fehr (1987) the anchor face was found to influence how the right target face was perceived emotionally. Reprinted from: Russell & Fehr, 1987

In a different study, Russell and Fehr (1987) investigated the influence of a second face presented along a target face. In a series of experiments they repeatedly found that the emotion was rated differently dependent on what type of face was rated before the target face. The procedure was comparable in all experiments: one, or later two, anchor faces with a distinct emotional quality were placed to the left of a target face, see for an example Figure 2. Participants were asked to rate the emotional quality of the left picture(s) first (the anchor face(s)), followed by the right picture (the target face). Changing the anchor picture resulted in a shift in the reported emotions for the target face, even when this target face expressed some degree of emotion in itself. When, for example, a positive anchor face was placed aside a neutral target face, this target face was perceived as relatively negative when compared to a presentation of a neutral face in isolation. Ekman and O'Sullivan (1988) criticised that the

BOX 3: VOCALIZED EXPRESSION OF EMOTION IN SPEECH

As stated in the introduction, it is generally assumed that a person's emotion is, consciously or unconsciously, communicated in speech (see for an overview Scherer, 2003). This assumption has been confirmed in various experiments with above chance recognition for most basic emotions (see Scherer, 2003). In these experiments participants were typically exposed to recorded, lexically neutral, utterances and had to categorize the utterance as belonging to one of a limited set of basic emotions, usually around seven (e.g. stress, anger, fear, sadness, joy, boredom and neutral). Scherer (2003, citing Scherer, 2001) reports an average accuracy for emotion recognition of 62%. In comparison, 78% of the time stills of facial expressions are correctly recognized.

Voices can be described in terms of their acoustical characteristics. Average pitch height, intensity and variability are typical examples of such measures. If these acoustic patterns of emotional voices are analysed in terms of correlation with judged, or intended, emotions, typical patterns of such emotional voice can be explored. A compilation of several of these factor analyses results is presented in Table .

Ultimately, knowledge of such correlations could allow for the prediction of emotional content, even when little is known about the origin of the sample. If, for example, an utterance features an increased intensity, a rise of the fundamental frequency (f_0), and a general sentence contour that is descending, this could be recognized as speech uttered by a person who experiences anger (compare with Table)

Table 1.
Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions (reprinted from Scherer 2003, based on Johnstone and Scherer 2000).

	Stress	**Anger/ rage**	**Fear/ panic**	**Sadness**	**Joy/ elation**	**Boredom**
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

general effect found by Russell and Fehr is only strong when neutral or ambiguous faces are used. Although there might be some truth in their argument, it does not seem to have much ecological value. First of all, emotional expressions are ambiguous, as revealed by, for example, the non-perfect recognition scores which were quite probably already based on rather prototypically acted material. Somewhat related, Scherer (2003) hints at an issue with experiments using stills of images: normally we only perceive moving faces. Most of what we know about recognition scores, however, has mainly been obtained using stills (p. 236). Second, more in relation to this thesis' topic, is speech somewhat more ambiguous than still imagery, and even that material has mainly been obtained using acted emotions (Scherer,

2003). Cowie and Cornelius (2003) suggest that emotional qualities in everyday speech may not be so strong at all; maybe speech prosody merely hints at underlying moods.

The experimental results of Cauldwell (2000) and Russell & Fehr (1987) seem to confirm that perceived emotion is influenced by the context in which it is perceived. Additionally, the environment is considered as one of the crucial descriptors, and discriminators, of emotional experiences (Barrett, Mesquita, Ochsner & Gross, 2007; Scherer & Zentner, 2001; see also Box 1). Altogether this leads to the suggestion that environmental audio may also influence the perceived emotion in speech:

H1: An auditory environment with a distinct emotional quality influences the perceived emotional quality of emotional speech

In the commotion model, discussed earlier, a second route was recognized as well: an alternative route in which the environment influences the prosodic features of the voice. This route is what will be discussed in the next sub-section.

1.3 Environmental influences on the sender

In subsection 1.1, "Communication of Emotion", it was mentioned that emotions lead to a change in voice prosody. Not only emotional experiences may give rise to such changes in voice, stress may also lead to comparable changes. Three main influences of the environment on the sender have been recognized (Laukka 2004). An environment may require a person to speak a) louder because of noise, b) softer because of norms, and c) different because of emotional affection. When the environment changes from silent to more noisy the sending person will probably have to speak louder above a certain level of intensity, an effect known as the Lombard effect. It has been noted (Junqua, 1996), however, that the Lombard effect varies not only with intensity, but also with different types of noise, e.g., babble noise vs. white noise, which may relate to the other two influences. Second, the sending person may be either restricted or unrestricted as a function of how appropriate it is to speak loud or soft in a certain environment (Picard, 1997; Laukka, 2004). This relates to some extent to the purpose of the conversation and the other persons in that environment (Brown & Fraser, 1979). Thirdly, the sending person may be emotionally affected by the environment, resulting in a mood change (Baber & Noyes, 1996; Junqua, 1996; Brown & Fraser, 1979; Scherer & Zentner, 2001).

1 A definition that resulted from ESCA-NATO "Speech Under Stress" workshop (Lisbon, Portugal,

BOX 4: STRESS AND THE LOMBARD EFFECT

Noisy environments induce a modification of the speech production, an effect that is known as the Lombard effect (see Junqua, 1996). This effect is described quite often, but mainly in the context of intelligibility under stressful conditions, e.g., speech recognition in cockpits.

The Lombard effect is often considered as an effect of stress, which influences voice prosody. Stress in speech, however, is rather ill defined. The decomposition of stress by Murray, Baber & South (1996), defining stress as the "observable variability in certain speech features due to a response to stressors" (p. 5) [1], allows for a better understanding of stress and thus understanding of the Lombard effect. Murray et al. discern four levels of stress (between brackets the stressor type description from Steeneken & Hansen, 1999): zero-order, external physical events act directly on the vocal tract, e.g. a trembling floor (physical); first-order, internal physiological effects due to, e.g., chemicals or lack of sleep (physiological); second-order, changes due to conscious mediation, e.g., trying to make yourself better understandable in noisy situations (perceptual), and; third-order, additional effects, which entails the more conscious control, including suppression, or compensation of lower order effects.

Murray, Baber & South (1996) assume that the Lombard effect is primarily caused by a perceptual stressor, and can thus be understood as a second order effect. When responding to external feedback, however, e.g., when responding to "Could you speak up a bit?", it is considered a third order effect, caused by a psychological stressor. Junqua (1996), however, notes that the effect might be "governed by the desire to obtain intelligible communication" (p. 15) in the first place. This corresponds with Murray et al.'s remark that the Lombard effect encompasses compensation for both noise at the sender's side, as well as (assumed) noise at the receiver's side. The latter reaction, however, is may be a learned one since observed children do not seem to compensate for noise at the receiver's side (Murray et al., 1996, p. 10).

The Lombard effect is not straightforward to simulate, aside from issues that were discussed above, the change in the voice depends on the nature of the speaker, the context and the environment. For example, the Lombard effect is different for two different types of sound, e.g., white noise vs. babble noise. Considering the Lombard effect as a binary (or a single dimensional) effect would be an oversimplification (Junqua, 1996).

One effect of noisy environments on the prosodic features of speech is widely known as the Lombard effect, see box 4. It is interesting to see how the changes in acoustic properties of Lombard speech seem to overlap with angry speech, but also that of joy. Junqua (1996) reports for Lombard speech an increase in the fundamental frequency, a shift in energy from low frequency bands to higher bands, and an increase of the overall intensity level as the main acoustic changes (compare this with Table in box 2). Cowie and Cornelius (2003) also hinted at this potential confusion: "A system for detecting stress by voice needs to take account of the possibility that emotions such as happiness may give many similar signs." (p. 12) It remains an open question, however, whether the changes in the acoustical features actually confuse human listeners, since there may be other features that have not been considered, such as taking into account the fact that the person is standing in a noisy environment.

September, 1995)

Within the commotion model, reasoning about the cause of something is part of the induction route. When someone is screaming in a loud and noisy environment, it may be most reasonable to assume that this person is screaming to make him or herself heard. This type of understanding, however, may change when this noise context is removed from the signal available to the receiver. Based on the comparison of the changes in speech prosody mentioned by Junqua (1996) and the remark of Cowie and Cornelius, it is thus expected that:

H2. Speech recorded in noisy environments is perceived less neutral when listened to outside the noisy context ; i.e. more negative and more aroused

Remember that the hypothesis does not state that Lombard speech is neutral. Lazarus (1999, p. 35, via Cowie & Cornelius, 2003, p. 11) points out: “when there is stress there are emotions (...) when there are emotions, even positively toned ones, there is often stress too.” Maybe, stressed speech is closely related to emotionally negative and aroused speech; maybe detecting a negative and aroused emotion in Lombard speech is actually a genuine emotional experience.

1.4 Relevance in product development

Aside from a purely academic interest, knowing about the relevance of environmental sound is important in communication technology. Improving the signal to noise ratio has lead to the development of several noise suppression techniques that allow for reduction of all sounds, with the exception of those sounds coming from the person speaking. It is, for example, possible to suppress the noise of a party, the music and the other talking people, when making a phone call from a party. The more traditional means of communication still transfer most of the sounds of events occurring at the sender's side of the communication line. Even though for quite some time noise suppression mechanisms have been in place suppressing stationary noise, sounds from unpredictable events are not filtered out. The technology, however, is progressing at a fast rate. Although the signal noise ratio's of the sounds are improved, one may wonder whether removing the environmental sound from this communication process does not alter how the speaker is being perceived. It may be quite confronting to a sender if he or she, without knowing, is perceived as an angry screaming person, whereas the only intent of this person was to make him or herself better understood at a party.

1.5 Outline of the thesis

This thesis is set out to test the following hypotheses:

1. An auditory environment with a distinct emotional quality influences the perceived emotional quality of emotional speech
2. Speech recorded in noisy environments is perceived less neutral when listened to outside the noisy context; i.e. more negative and more aroused

Before we can test the two hypotheses, however, material for the experiment needs to be generated. Especially when it comes to emotional speech obtaining this material is a non-trivial task. Therefore two experiments described in this report are dedicated to generation (Experiment 1) and selection (Experiment 2) of emotional stimuli for Experiment 3, which is set up to verify the first hypothesis. In the last experiment, Experiment 4, Lombard speech will be addressed to verify the second hypothesis. With a set up comparable to experiment 3, Lombard speech will be played randomly with different intensities of noise. This report will end with a global discussion of the obtained data and give recommendations for both theoretical and practical future work.

2. Experiment 1: Collection of emotional and neutral speech

2.1 Introduction

Before the effect of environmental audio on the perception of speech can be studied, a database of stimuli is required. Ekman and O'Sullivan (1988) commented on the claims by Russel & Fehr (1987) that relativity in perceived emotion only affects relatively neutral stimuli, or emotionally weak utterances. To investigate whether this also applies to possible relativity in the emotional perception of speech, also utterances expressed in an emotional way are required, making creation of the database a less trivial task.

Despite the numerous studies on emotion perception, none of the available databases of emotional speech suited the needs of what has become Experiment 3. Either the language was other than Dutch, the quality too low, or too little was known about how the stimuli were generated. Hence it was decided to create a new stimulus set for this experiment. Instead of recording only sound, also video was recorded for potential use in other experiments.

After considering several options of creating emotional stimuli, among those mentioned by Scherer (2003) and Westermann, Spies, Stahl and Hesse (1996), and personal communication with Krahmer and Swerts (March, 2008; see also box 5), it was decided that a mood induction procedure (MIP) was the most appropriate generation method. Potential alternative methods to create a set of emotional speech stimuli were thought to be less suitable for the following reasons: Acted speech often results in, although very prototypical, overacted speech, thereby loosing ecological validity. A database of natural stimuli, suitable for experimentation is hard to create, and the original emotional intent is difficult to control. Last, synthesized speech still sounds too artificial to be convincing.

The type of material required is speech material, hence the Velten mood induction procedure (or Velten MIP; Velten, 1968; Kenealy, 1986) seemed an appropriate method for inducing emotion. The Velten method consists of increasingly more emotional sentences that the participant needs to read out aloud. Starting, for example, rather neutrally with the sentence "Today is neither better, nor worse than yesterday", sentences become increasingly more emotional, ending with, in case of the negative Velten MIP: "I want to go to sleep, and

BOX 5: DATABASES OF EMOTIONAL SPEECH

Experimental research requires material which varies only in predefined variables, preferably only the variable of interest in the experiment. In emotion research, obtaining such material is hard. While one could try to gather real life speech samples, it is difficult to obtain a high quality material. On the other hand, artificially obtained samples may lack ecological validity.

Four main categories of databases can be discerned, based on the method of generation: databases based on natural vocal expression, induced emotional expression, simulated emotional expression and synthesized emotional expression (Scherer, 2003).

Natural vocal expression has as major advantage its high ecological validity, but for several reasons it is less suited for experimentation (Scherer, 2003): sets are often limited to a small number of speakers; recording quality is often suboptimal and it is hard to determine the precise nature of underlying emotion. Additionally, it may be hard to obtain lexically neutral content. A way to resolve these problems is by inducing emotions.

Emotions can be induced either directly using, e.g., drugs or indirectly by putting speakers under considerable stress, playing back emotion-inducing films, music or slides, etc. It is noted by Scherer (2003) that the effects of these procedures often produce only weak effects and it is not possible to assume that the emotional states of all persons are similar.

An third method of generating emotional stimuli is by asking humans to act or portray emotions vocally. It is suggested that lay-actors may not be ideal in this type of stimulus generation since they are unexperienced with expressing convincing emotions (Scherer, 2003; and see e.g., Burkhardt, Paeschke, Rolfes, Sendlmeier & Weiss (2005) for an example). Emotions expressed by actors, however, are sometimes thought to be stereotypical (Scherer, 2003; Burkhardt et al., 2005) and lack ecological validity. Scherer (2003), however, counter argues that maybe all publicly observable expressions are to some extent portrayals, and since they are reliably recognized, they probably reflect at least some 'normal' expression patterns. Still, the emotional expression may be a learned emotional expression; artificial, but agreed upon within a certain culture.

The last alternative type of generating emotional material is by synthesizing emotions either by synthesizing voice and its parameters from scratch or by re-synthesising (e.g., rearranging the fundamental frequency) original material (Scherer, 1995). The variable properties in synthesized emotional speech may have been derived from earlier research using the other methods listed. The main disadvantage of this method is that it introduces many unwanted artefacts.

Although Scherer (2003) seems to prefer the stronger, more prototypical expression of acted emotions, over the relatively weak emotional expression resulting from emotion induction (for an actual comparison, see Krahmer & Swerts, in press), it may be that the less prototypical emotions are actually more like natural emotions. Cowie & Cornelius (2003) openly question whether normal speech is strongly affective, after having reviewed the Belfast Database of emotional speech, which contains many real world speech samples, “emotive topics produced very little speech suggesting anything approaching full-blown emotion” (p. 9). However, while “speech rarely expresses strong, pure emotion, it is not often emotionally neutral either (...), 84% of cases where a clip was rated neutral included a supplementary label indicating that some emotion was felt to be present.” (p. 11).

never wake up”. Advantage of this induction method is that participants are already reading sentences aloud for the induction procedure. The drawback, however, of the Velten method is that it uses sentences that are deliberately emotional. Since users have to assess the emotion purely on prosodic features, it is not desirable to have lexically emotional sentences. One may thus wonder whether the Velten MIP is in that case that much different from the usage of,

e.g., film, or music, as induction stimulus. The Velten MIP was still thought to be interesting because it includes already some 'preparing' of the voice during the actual procedure. To increase the chance, however, that the sessions with the participants resulted in usable emotional utterances, a second induction method was chosen as well, namely the film-mood induction procedure. Emotion elicitation using emotional film fragments is interesting because of its relatively high effectiveness in inducing emotions (Westermann, Spies, Stahl & Hesse, 1996). Additionally, it was hypothesized that a combination of the two methods might even result in a stronger emotional experience.

Some authors have criticised the Velten MIP for its demand characteristics; the Velten method often implies a request to the subjects to act as if they experience the emotional content of the sentence (Westermann, Spies, Stahl & Hesse, 1996). In the suggested procedure, however, this request to enact the emotion of an utterance, does not really apply when the participant is required to read out aloud a neutral sentence. It is, however, not possible to rule out that all prosodic features are just part of the speech due to a short term habituation to the enacted sentences, not because some real emotion was felt. This problem, however, does not play a role with the second induction method using films.

2.2 Method

Design

The experiment was designed as a two treatment pre-test/post-test design, repeated twice. In the first trial, induction of a neutral emotion took place, followed in the second trial by induction of either a positive or negative emotion.

Within each trial, two treatments took place, first an induction using the film MIP, followed by a Velten MIP. After each treatment participants were required to utter a lexically neutral sentence, which was expected to be recognizable as emotional or neutral speech depending on the intended emotion.

Before, in between, and after the treatments, the participant's mood was assessed through a method of self report, using eight seven-point bipolar scales that corresponded to the Valence-Arousal model of, e.g., Russell (Russell, 1980; Yik, Russell, Barrett, 1999)). The scales covered both the valence and arousal axes and a 45° rotation of these two dimensions (Yik et al., 1999), see also Figure 3.

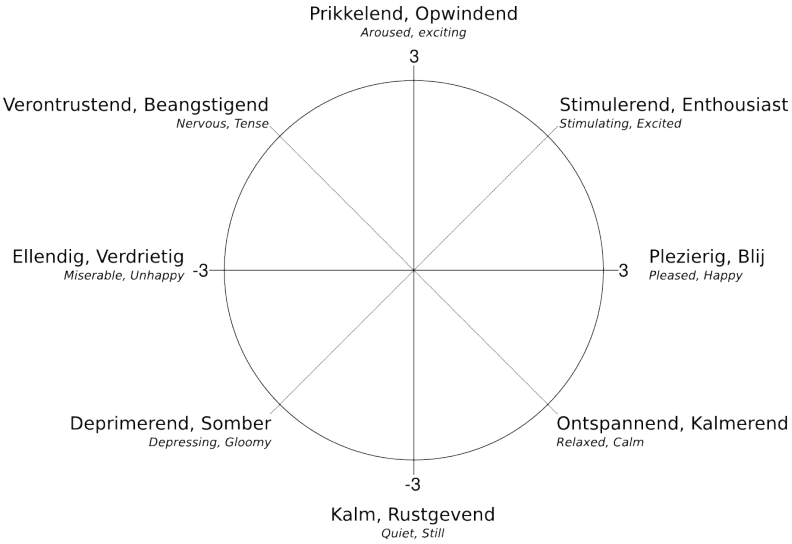

Figure 3. Relations of several emotions within the valence-arousal model. Adapted from Yik, Russell, Barrett (1999)

Participants

In this experiment mainly interns, students, at Philips Research participated during their regular working hours. Three participants participated in reply to a post on the student pool mailing list of the Tilburg University. These students got course credits for their participation. In total, 17 persons participated, all born in The Netherlands. Their average age was 25.3 years. Nine were male, eight female. All participants were naive about the goal of the experiment and were told that they had to rate the emotional impact of several products made by Philips Design when influenced by emotion. This fake purpose was made explicit to the user to help the participants with concentrating on the induction procedure (Westermann, Spies, Stahl, Hesse, 1996). Talking aloud is part of the Velten procedure, and was explained to the participants as such. The camera was explained to the participant as necessary for 'further analysis'. The real reason, however, was potential use of the utterances in other research. The participants were randomly assigned to an experimental condition by shuffling the feedback forms that had a non-informative letter (A and C were positive, B and D negative). All the participants gave full consent for usage of the obtained material in research and publications.

Apparatus

Film induction requires relatively short fragments, for practical reasons, featuring strong emotional content. Although it is desirable in film induction to use a standardized set

BOX 6: MEASURING EMOTION

Humans experience something they call emotions, but how to measure it? Emotions can not simply be measured like distance with a ruler. It is an experience still only clear to the person experiencing it. Like for example the experience of loudness, but the advantage of loudness is that there is a close relationship with something changing in the physical world, e.g. the change in intensity of the sound. There is no such equivalent for emotions. And that makes measuring emotions difficult, therefore the method of measurement is still much debated. In this box a short overview will be given.

There are three classes of emotion measurement: self-report, physiological and interaction based. Self-report measures are most popular, most likely because the accessibility of this measure; they are relatively easy to assess. A sheet, or a set of sheets, and a pencil is often enough. A division in this measurement type can be made on the basis of the theory it is based upon, e.g. discrete or dimensional. Although both types allow for the use of scales, dimensional theories assume that it is possible to encode emotion as a two or three dimensional entity, whereas discrete emotion theorists believe that this thought is too simplistic, and refrain from coupling certain emotional experiences. In larger, excessive questionnaires, the differences may be not so apparent; dimensionalists will ask ratings on different emotions and will convert them to a position in the n-dimensional model, whereas discretionists will also allow for rating of discrete emotions. When quick assessments are required, however, a dimensionalist is most likely to present 2 or 3 (positive-negative, sleepy-aroused) scales, whereas a discretionalist is more likely to present a limited number of labels, from which the user has to choose (neutral, anger, fear, joy/hapiness, sadness, disgust surprise). Both methods have their obvious limitations. The dimensional approach is not always able to discern between quite distinct emotions (e.g. anger and fear are both in the negative-aroused quadrant). This can partly be resolved by adding a third dimension (e.g. amount of control). Discrete emotions, on the other hand, may be quite straightforward with low numbers through exclusion of options. But as long as participants do not score all stimuli correctly all of the time, the degree of confusion may still be informative.

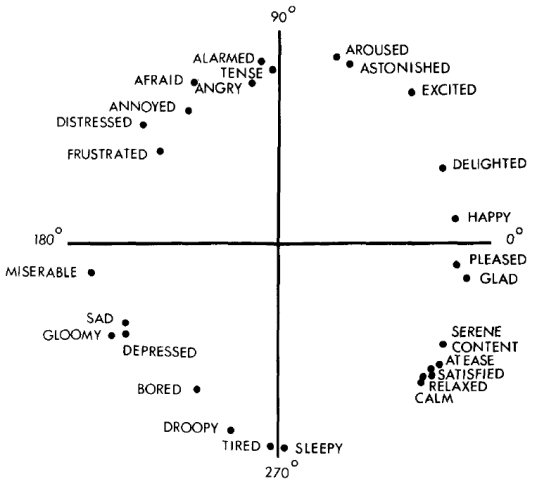

Figure 4 . A typical example of a dimensional organisation of emotions (Russell, 1980) .

The second class of emotion measurement is using physiological measures. Typical measures used are hart rate and skin conductance. While changes in these physical properties may occur due to a change in felt emotion, the results are not always reproducible and stable (e.g. Barrett, 2006). Additionally, it is hard to directly infer from these properties that emotions are felt. It is, for example, possible that the increased heart rate is caused by the fact that the person has been running some time before the measurement.

An alternative class of emotion measurement is less concerned with quantification. To the contrary, as emotions in this model are considered as interaction: “dynamic, culturally mediated, and socially constructed and experienced” (Boehner, DePaula, Dourish & Sengers, 2007, p. 275). The observation of how users interact with objects, how they respond in conversations to others, the words they use etc., is used as input to a more qualitative analysis of the user's emotional experience. In addition to observations, (semi-structured) interviews, assisted with emotional foils, etc. can be used. While less, or not, suitable for quantifiable data, it may be an interesting tool in evaluation processes.

of emotional material (Rottenberg, Ray & Gross, 2007), much of the suggested material by Rottenberg, Ray and Gross (2007) was not suitable for our purposes. The positive and negative stimuli recommended were either not experienced as exceptionally positive (Fake orgasm from the “When Harry Met Sally” movie), quite funny, but not available with proper Dutch subtitles (“Robin Williams Live”), too simplistic (“The Champ” dying scene), or only available as a low resolution clip (the neutral 'Sticks' screensaver). Krahmer (personal communication, March 2008) suggested other fragments that he had used succesfully in earlier experiments involving mood induction (e.g. Krahmer, Dorst & Ummelen, 2004). In the positive condition, a 5 minute fragment from the Friends series was used, a popular sitcom. It was the edited start of the episode “One where everybody finds out”, focussing mainly on the discovery of the secret relationship of Monica and Chandler by Phoebe. In the negative condition, a 7 minute fragment from the film “Schindler's List” was used in which Jewish families are violently forced out of their houses in Krakow. In the neutral condition a similar length fragment from Animal Encounters was shown, an informative nature program from the Animal Planet Network, featuring underwater scenerey from the Great Barrier Reef.

Sentence material for the Velten method was obtained from Wilting (2005), who translated the original sentences of Velten (1968) to Dutch and showed that the translated material was effective in changing the emotional state of naïve participants (see Appendix A for the selected sentences).

The goal of this procedure was to generate lexically neutral, but vocally emotional, stimuli. Although one lexically neutral target sentence could have been enough, four different sentences were recorded, simply to increase the chances of obtaining proper material. Two sentences were recorded just after the film induction, and the other two just after the Velten induction. Two of these sentences were sentences obtained from earlier emotion research in the Netherlands (e.g. from Mozziconacci, 2001; de Gelder & Vroomen, 2000):

S1. Zij hebben een nieuwe auto gekocht (They have bought a new car)

S2. Zijn vriendin kwam met het vliegtuig (His girlfriend came by plane)

It was thought, however, that these sentences were too concrete and detailed; more ambiguous sentences might allow for a larger variation in different contexts. For this reason also two alternative sentences were produced. Although these sentences maybe less neutral, they are expected to be more susceptible to alternative emotional explanations. This idea

followed from Cauldwell's (2000) finding with also a quite ambiguous sentence (“What do you mean?”)

A1. En dat was dus de reden (So that was why)

A2. Zometeen ga ik dat nog halen (I will get that in a minute)

The first sentences of both pairs (S1 & A1) were recorded after the Film induction, the second sentences (S2 & A2) after the Velten induction.

Material was recorded at an empty office within the Philips research department using a AKG CK 31 cardoide microphone mounted to a AKG LM 3 connector. The signal was pre-amplified using a RME QuadMic preamp. The two channel signal (XLR) was wired to the Sony HDR-SD7 video camera which converted the analogue signal to digital, and was also responsible for recording the participant's face. The audio was encoded in 48kHz AAC at 448 kb/s (for 5:1 audio). The visual material was of no relevance to this research in particular, but was recorded to the benefit of other researchers. The two channel XLR signal was extracted and converted back to single channel audio using Audacity 1.2.6.

Figure 5 . The frontal view of the set-up used in the induction experiment, as seen by the participant

Instructions and film fragments were presented on a 17" wide screen monitor (Philips 17PF9945/37) which was positioned at about one meter distance from the participant. The instructions itself were prepared using Microsoft PowerPoint 2003, with automatic slide progression where possible. The delay between each sentence was fixed to 20 seconds.

In order to assess the mood, a pen, and a package of sheets was prepared on which the participant could fill in personal details, assess his or her mood during the experiment, and report the product evaluations. The package was constructed such that when adhering to the request to always turn the page after completion of the self report, the participant was unable to check his or her earlier mood reports.

Participants were initially told that the purpose of this experiment was to study the influence of emotions on how Philips products are perceived. For this fake experiment, product photos were downloaded from the Philips Design iF Design Awards 2008[2] page.

TightVNC[3] was used to monitor the participant's screen remotely, giving the experiment leader also the ability to control the computer when problems would occur, even though no such problems did occur during the actual experiments.

Procedure

Participants participated one at a time, in a small conference room within Philips Research. After a short welcome, participants were asked to sit down and read everything presented at the screen, including the instructions, aloud. After instructions on how to operate the computer were given, and the basic outline of the experiment was presented, the instructor left the room.

The first task for the participant, that followed after the instructions, was a request to report his or her mood. This mood had to be reported on a sheet of paper provided to the participant using eight semantic differential scales (based on Yik, Russell, Barrett, 2003; see for details Appendix B),. When finished reporting, the participant was asked to turn the page on which the mood was assessed and the participants could start the first condition using the space bar. Note that there was only one mood assessment report possible on each page. A neutral video fragment started, after which the first two target sentences (A1 & S1) were shown directly after a short slide remembering the participant to read out aloud everything. Again, the user was requested to report his or her mood using eight bipolar scales. When finished the participant could start the neutral Velten MIP by pressing the space bar. Seven sentences (see Appendix A) were displayed for 20 seconds. During these 20 seconds, the participant had to read the sentence in silence, after which he or she had to read the sentence

2 Philips Design iF Design Awards - http://www.design.philips.com/about/design/designawards/ifawards2008/
3 TightVNC: VNC-based Free Remote Control Solution - http://www.tightvnc.org/

aloud while the sentence was on display. The participant was asked first to read the sentences in silence, a The last two of these seven sentences were the other two target sentences (A2 & S2). After presentation of the last sentence, the user was again requested to report his or her mood. The first part was concluded with a sequential display of 10 Philips products, which the user had to rate on a valence and arousal scale.

After rating the product photo's, participants were given the opportunity to take a short break. The second trial followed the scheme of the first part, but instead of the neutral film fragment and the neutral Velten sentences an emotional film fragment was shown and emotional Velten sentences were used. The Velten MIP consisted in this condition of 25 emotional sentences plus two lexically neutral target sentences (see Appendix A). After finishing the ratings of ten other Philips products, the user was informed that the experiment had ended and the experimenter could enter any minute. The total experiment lasted about 45 minutes.

2.3 Results

Valence and arousal was measured using eight semantic differential scales. Internal consistency with the measures valence and arousal was checked by calculating the Cronbach's alpha. The Cronbach's alpha of both the Valence scale ($\alpha = .89$, nr. of items=6) and Arousal scale ($\alpha = .91$; nr. of items=4) were found to be good (see Appendix B for more information).

Figure 6 shows how the mean self reported emotion changed during the various stages in the experiment in two dimensions, arousal and valence. The error bars represent the 95% confidence intervals for the mean. Induction of the emotion can be considered successful, as the valence scores for two groups clearly diverge after exposure to the emotional film fragment. Since the goal of the induction procedure used was to influence the valence, only the effect on self reported valence will be discussed in more detail.

Although on average, participants felt rather positive when they started, the neutralisation using the neutral film fragments and Velten method was successful; the difference between the start condition and the end of the neutral emotion induction was significant, $t(17)=4.00$, $p < 0.001$, $r = .97$. It should be noted, however, that the reported valence is still relatively positive, as the confidence intervals do not cover the neutral .0 point, see Figure 6 . The effect of the positive induction was not significant when the start of the positive induction trial and the end of this trial, after Velten, are compared, $t(8)=1.02$, *ns* , $r = .25$. From the confidence intervals in Figure 6 , however, one can see that the valence as reported after the Film induction was higher. In both cases, the reported valence is significantly above the neutral .0 point, and thus positive. Last, the effect of the negative induction was significant; the valence at the end of the negative induction procedure was significantly different from the valence as reported at the start of the negative emotion induction procedure, $t(9)=3.84$, $p < 0.00$, $r = .93$. The value is also significantly below the .0 neutral point, indicating a clear negative emotion.

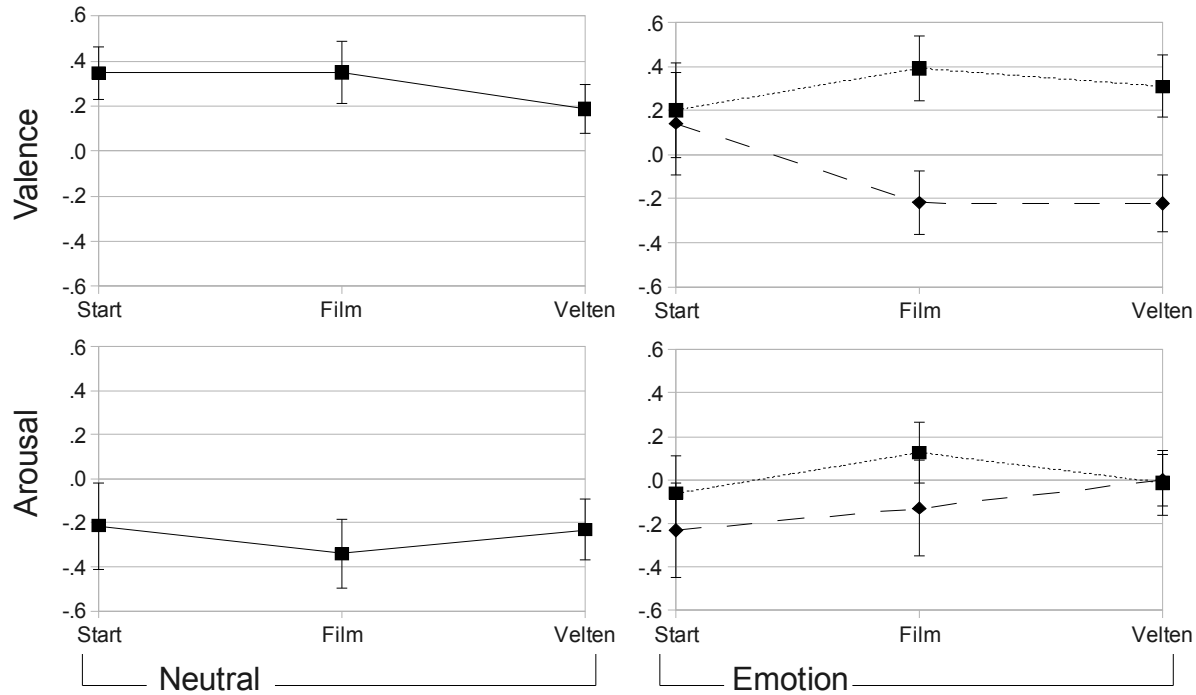

Figure 6 . Change of the average valence and arousal scores over time. The solid line covers both groups, whereas the negative and positive groups are represented with a dashed line and a dotted line, respectively.

2.4 Discussion

The goal of this experiment was to generate material for the stimulus set that is required for Experiment 3. It was demonstrated that the (cumulative) induction procedure was able to influence the reported mood. Obviously, it was not an experiment capable of

evaluating the two mood induction procedures (MIPs) as there was no randomized order of the different procedures.

The self reports showed a significant difference in valence between the two groups after induction using the film and Velten methods. Which was the result that was expected: the two mood induction procedures were expected to influence the participant's mood, either towards a more positive mood, or towards a more negative mood. The film based induction procedure seemed quite strong on its own already in influencing the participant's emotion. It is not possible, however, to tell whether the Velten method on its own could have produced a similar effect, due to the lack of randomization of the two induction methods.

Whether the reported mood is also the actual mood experienced by the participant can not be guaranteed. The change may be due to a demand characteristic; participants may have thought that reporting a more emotional mood after the first trial was desired even though it was not actually felt. There is, however, little reason to assume that this has influenced the way the target sentences have been uttered, as none of the participants were able to guess correctly what the role of these neutral sentences was in the experiment. A sentence uttered after the Velten MIP, however, may be influenced by a possibly unconscious, but also unemotional habituation process; participants may have 'stuck' to the tone of voice. Despite this possibility, the extensive successful usage of the Velten MIP in other experiments (Kenealy, 1986), signals that it is quite safe to assume that emotions are infected, and hence at least some of the change in voice prosody is caused by a change in the felt mood.

It is not warranted that a change in voice prosody did occur. If it does, it is, given the observations in the previous paragraph, safe to assume that when changes are heard in correspondence to the induced emotion in the voice's prosody at least some genuine emotions were felt while uttering the sentences, especially when emotion is heard in utterances uttered after the film MIP. To assess this, emotions should be randomly and blindly be presented to a set of listeners, which is what the next experiment was designed for.

3. Experiment 2: Evaluating the speech samples

3.1 Introduction

Experiment 1 resulted in over 100 samples, and not all of them were expected to feature the characteristics sought after, namely relatively clear 'positive', 'negative' and 'neutral' speech. Additionally, a factorial design was planned for Experiment 3, in which utterances recorded in isolation would be randomly combined with multiple environmental sounds. A factorial design with 100 samples, however, would result in an unmanageable amount of conditions. The goal of Experiment 2 was to make a good selection of samples, some of which represent utterances expressed in an emotional way. Such emotional utterances were required to investigate the comments made by Ekman and O'Sullivan (1988) on the research by Russel & Fehr (1987) that a emotional expression may be a more stable percept, when compared to an emotionally neutral expression.

Although it would have been possible to make a selection of samples based on the self-report measures of the participants, it would not have warranted that the selected samples actually featured the information that communicates the emotion. Additionally, large in-between person differences were expected in self-report of the different speakers. Ideally, however, self reported emotion and the strength of the heard emotion correlate reasonably, given the assumption that emotions can be heard in speech.

3.2 Method

Design

The independent variable in this experiment was the speech sample presented, the dependent variable was the rated emotion on the dimensions valence (positive-negative) and arousal (aroused-sleepy). In total 104 samples (from 13 different persons, 8 utterances per person) had to be rated which were randomly presented to each participant[4]. Remember that for each person neutral sentences had been recorded, but that only half of the participants were exposed to positive induction stimuli, whereas the other half were exposed to only the negative stimuli. As a result, there were as many neutral samples as emotional utterances.

4 In the previous experiment 17 persons participated. Recording of the audio, however, failed for four participants due to a broken cable between microphone and camera

Participants

In total 8 participants participated, but not everyone rated all utterances. On average each fragment was rated by 5.84 participants; range 5-7. Four of the participants were male. Age was administered using age-groups (range 10 years). Most of the participants belonged to the age group of 21-30 years old. One participant was younger (11-20), and two were older (41-50 and 51-60).

Apparatus

An on-line questionnaire was prepared for rating of the utterances. The utterances were recordings made during the first experiment and encoded as 128 kbs MP3 at 44.1 kHz, 16 bit. The on-line questionnaire was programmed in PHP, and audio playback relied on a small Flash-based player[5]. All randomization was done by the built-in PHP 4 random function. Analysis of the utterances was done using Praat[6] 5.0.02, a free and open source speech analysis tool developed by Paul Boersma and David Weenink of the Institute of Phonetic Sciences (University of Amsterdam) and GIPOS 2.1, a speech analysis tool developed in the nineties at the former Instituut voor Perceptie Onderzoek (IPO) in Eindhoven.

Procedure

Participants were approached via e-mail and/or direct communication. They were asked to visit the website that could be accessed using the URL provided. The goal of the experiment was explained at the first welcome page, basically asking the potential participant to rate a series of speech samples' emotional content. In order to continue, users were asked to fill in some personal details. By pressing the next-button, the participants were presented with a short audio test, testing whether volume settings were all right, and the browser supported Adobe Flash and JavaScript. After confirming that everything worked as expected, the participant was presented with a random sample. The participant was asked to judge the emotion of the person talking using two 7-point bipolar scales (negative-positive, passive-active), ranging from -3 to +3. For each label three synonyms were suggested to make the possibly abstract extremes more clear. For the positive (positief) label these synonyms were: happy (blij), satisfied (tevreden), fortunate (gelukkig). For the negative (negatief): sad (verdrietig), unfortunate (ongelukkig), disappointed (teleurgesteld). For active (actief):

5 XSPF Web Music Player (Flash) - http://musicplayer.sourceforge.net/
6 Praat: doing phonetics by computer - http://www.fon.hum.uva.nl/praat/

aroused (opgewekt), intense (intens), exciting (opgewonden) and, last, for the passive (passief): sleepy (slaperig), tired (moe), relaxed (ontspannend). The rating procedure was repeated until all samples were presented and rated once. Those who rated all samples (n=5), took an average of 27.18 minutes to complete the test.

3.3 Results

For each recorded utterance, at least 5 ratings were obtained on two scales, that of valence and arousal. The values were averaged and divided by three, resulting in values within the -1 to +1 range. Since similar values have also been obtained just after recording of the utterances in the previous experiment, it is possible to compare the averages obtained in this experiment with the self reported mood of the speaker at the time the utterance was recorded. Correlation between the self reported valence and arousal was only weak, r = 0.27. Among the 25 utterances with the lowest valence rating (valence range -0.83 to -0.28) there were 15 neutral utterances, nine negative, one positive. Three out of the nine negative utterances were uttered after Film induction. Among the 25 utterances with the highest valence ratings (valence range 0.22 to 0.87) there were ten neutral utterances, thirteen positive and two negative utterances. six out of the thirteen positive utterances were uttered after Film induction

Based on a combination of the results of this experiment and the intended emotion, and partly the neutrality of the same sentence by the same person but with intended neutral emotion, 4 positive and 4 negative samples were selected that were thought to be good representatives of each emotion. The selection was lead by the scores obtained in this experiment. An utterance was, however, disregarded when the intended emotion did not match with the perceived emotion; i.e. if a file with a distinct positive rated valence was recorded after a negative or neutral induction, or a file with a distinct rated negative valence was recorded after a positive and neutral induced, utterances were disregarded. Additionally, it was ensured that the neutral equivalent of the sentence was not among the top 25 most negative, or 25 most positive sentences. Along with the eight emotional utterances (four of which were positive, and four negative) selected, the eight neutral counterparts, uttered by the same speaker and using the same induction method, were selected as well.

Analysis of the speech samples

Reviewing the vocal communication in speech, Scherer (2003) presents a table

summarizing research describing correlations between emotions and the basic acoustic patterns in speech (see Table). Although the set of material obtained in the previous experiment is too small for a factor analysis[7] including all the properties identified by earlier researchers, it is possible to compare the typical changes as reported by Scherer, and the direction of change in the recorded stimuli selected. The results of this analysis can be found in Table 2.

After analysis of each sample, values for the neutral stimuli were subtracted from the values obtained for the emotional stimuli (within person). Hence, a negative value represents a decrease in the property relative to the neutral state, and a positive value an increase in the property relative to the neutral state. The reported value is the average change over the four participants from the neutral condition, and significance reported is the significance of the change (testing the null hypothesis that neutral and emotional speech has the same values)

Table 2.
*The typical changes in speech prosody as noted by Scherer (2003) compared with the results of the analysis of the samples obtained in this experiment. ** = p < .05, * = p < .10*

	Negative		**positive**	
	Present study (n=4)	**Scherer (2003) Sadness**	**Present study (n=4)**	**Scherer (2003) Joy/elation**
Intensity (dB)	+0.11	↘	+0.45	↗
F0 mean (semitones)	+6.31	↘	+1.17	↗
F0 stdev (semitones)	-0.95	↘	+1.50**	↗
F0 range (semitones)	-2.90	↘	+3.36*	↗
Contour trend (semitones/s)	-1.64	↘	-0.10	
Relative high frequency energy	-.01**	↘	+.01	(↗)
Speech/Articulation Rate (syl/s)	-0.25	↘	0.19	(↗)

All pitch related values are obtained from close copy stylisations of the pitch contour. These were made by an experienced close-copy stylist from the former IPO institute. The result of this close copy contour is a series of time/log-frequency value pairs describing the

7 Hair, Anderson, Tatham & Black (1995) recommend to have at least five times more observations than the number of variables for a solid factor analysis

overall pitch contour when straight lines are drawn in between (in the logarithmic domain). Average pitch is defined as the average pitch value in this close-copy contour. Standard deviation is the standard deviation of this pitch from the mean pitch, and range as the maximum and minimum frequency values reported. The contour trend is defined as the slope of the regression line through the close copy contour value pairs, which is considered as a good approach (Hermes, personal communication, May 2008). It should be noted that this method might deviate from the method used by Scherer whose method is unknown to the author of this thesis. Note that automatic pitch extraction algorithms do not detect the intended pitch of the speaker, either pitch is detected in unintentional and irregular vocalisations, or the pitch found is an octave or two higher or lower than the actual intended and heard pitch due to the limitations of the algorithm. Close copy stylisations are corrected for these issues.

Speech and articulation rate are defined as syllables per second. Normally, speech and articulation rate are discussed as two separate measures, however, none of the utterances recorded featured pauses; articulation rate is the number of syllables per seconds, not considering pauses, whereas speech rate is the number of syllables per second including pauses (Boves, 1984, p. 108, referring to Butcher (1981)).

Additionally, the measures intensity and relative high frequency energy are mentioned. Intensity is reported in decibels as reported by the algorithm as implemented in the package “praat”, and the relative high frequency energy, which is defined as the proportion of energy in the frequencies higher than 1000 Hz, relative to the total amount of energy contained in the sample.

3.4 Discussion

The main purpose of this experiment was to make a selection of utterances that featured recognisable emotional speech. A selection criterion could have been simply the self reported mood, but it was uncertain whether self reported mood entailed recognizable encoding of that mood or emotion. Therefore a small group of participants was asked to rate the emotional content. No hypothesis was tested in this experiment and given the small number of participants such would not have been possible either.

In the introduction to this experiment, it was mentioned that ideally the self reported valence and the perceived valence obtained in this experiment should match. A correlation

was found, but it was only a weak correlation. The result can be explained by looking at the difference in method, and the difference in design. The method used to assess the emotion was based on self felt emotion using eight scales, the participants of this experiment were only able to use two scales (negative-positive and passive-active) to assess the perceived emotion. But the main difference can probably be attributed to personal differences. The value obtained in this experiment was a mean value based on at least five judgements by participants who had a large set of samples to compare the utterance against. The participants of Experiment 1 had to assess their earlier experienced mood, and had nothing but their own mood as reference material, which might be a source for in between persons differences.

As an additional check, the prosodic features were analysed and statistics were compared to the changes reported in earlier studies. The analysis of the selected samples showed only limited correspondence with the predicted changes in prosodic features when emotional speech was compared to neutral speech; most effects were too small to be significant. There where changes were significant, however, the change was in correspondence with the predictions from literature.

The limited correspondence is not thought to be a bad property and probably relates, aside the low number of samples being analysed, to the method to generate the stimuli. Although it is unclear what material was used to create the summary of Scherer (2003), much of the earlier research was based on stimuli by actors. Acted emotional speech is quite likely to be relatively more expressive when compared to utterances by (non-)actors experiencing induced emotions. Acted emotional speech, however, has often been criticised for being too prototypical and overacted (Scherer, 2003; see also box 5). Given these observations, the utterances recorded may not be so bad after all.

Having a reasonable set of stimuli, an attempt can be made in investigating how these stimuli interact with different environmental sounds. This is what will be discussed in the next section.

4. Experiment 3: Perception experiment

4.1 Introduction

So far, all experiments were mainly preparatory experiments. The goal of these experiments was to generate and select utterances suitable for the experiment described in this section, designed to test the first hypothesis as stated in the introduction, namely that an auditory environment with an emotional quality influences the perceived emotional quality of emotional speech.

Based on the experiments by Russell and Fehr (1987), where an visual anchor induced a change in the perceived emotion in a target face[8], a more precise prediction may be made in relation to this experiment: an anchor stimulus, in this case the environmental sound, should displace the target stimulus, the neutral or emotional speech stimulus, within the two dimensional arousal-valence model space away from the position of the anchor, increasing the contrast between the anchor stimulus and the target stimulus when compared with the distance between both stimuli in isolation. This is illustrated in Figure 7.

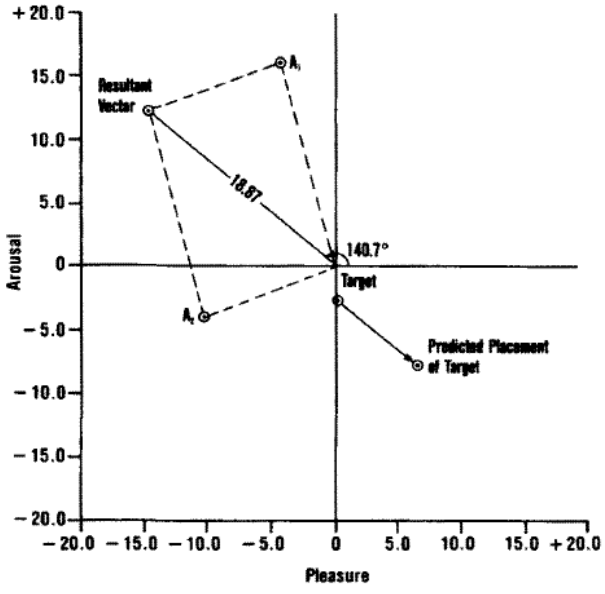

Figure 7 . Predicted displacement produced by a virtual anchor. The originally neutral target (in centre) is perceived differently when two anchors have been rated first. Displacement can be quite reliably predicted based on the virtual anchor. Reproduced from Russell and Fehr (1987)

8 The experiments by Russel & Fehr (1987) were discussed in more detail at page 5.

More intuitively, it was thought that a person talking in a room with uplifting, happy sounds, is not just feeling plain neutral. If this person is actually standing in this uplifting environment, and he or she sounds neutral, this person is probably relatively sad, otherwise his or her voice would at least inhibit some of the enthusiasm of the other people in the room, e.g., humans tend to laugh when they see other people laugh. For similar reasons, it is also expected that a happy person does not have to be exceptionally happy in this same uplifting, happy, environment. Conversely, it is also expected that a neutrally speaking person in a sad/ misery environment has been able to 'distance' him or herself from the environment, feeling relatively good.

On the other hand, there is also the result of Cauldwell (2000), who found quite the opposite result in his experiment. The perceived emotion of the voice was perceived as less angry when told about the relative peacefulness of the context. Maybe the influence of context is more complex than initially thought, and therefore the hypothesis is limited to a prediction of change, not the direction of change.

4.2 Method

Design

The design used in this experiment was a full factorial design in which both environmental sounds and emotions were varied as the independent variables. The five environmental sound conditions were: no sound, positive environment, negative environment, noise of medium level and loud noise. The emotions were represented by 16 utterances in total, those selected in Experiment 2. Four of these utterances were expressing a positive emotion, four a negative emotion and eight utterances were neutral utterances, which were, except for the emotion, equivalent, i.e., same speaker, same type of sentence (standardized or ambiguous) and same mood induction procedure used (film or Velten). Additionally, 4 other sentences were added to the test and presented to the participant, but these were only of relevance to Experiment 4.

The dependent variable was the rated emotion of the utterance played in context. Rating of the utterances was done using two 7-point bipolar scales for valence (negative-positive) and arousal (active-passive), comparable to how emotion was assessed in Experiment 2. Ratings of the environment was done using the same procedure as in Experiment 1: each environment was rated twice by each participant using eight 7-point bipolar scales.

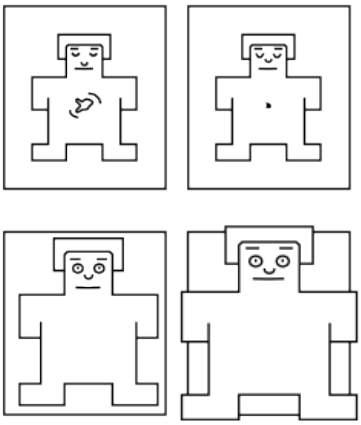

Figure 8 . An example of a graphical emotion assessment method, SAM (Bradley & Lang, 1994)

Method of measurement . In the previous experiment, not much attention was paid to how to assess emotion quickly. Assessment using eight semantic differential scales for every combination in the factorial design was found to be annoying by participants in pilot tests. Several quick assessment techniques were considered for use in our experiment, and special interest went into graphical methods such as the Self Assessment Mannequin (SAM; Bradley & Lang, 1994; see Figure 8) and the Product Emotion Measurement Tool (PrEm; Desmet, Hekkert, Jacobs, 2000; Desmet, 2003). The problem with these methods, however, is that they directly link emotion to an, somewhat iconified, facial expression. It was thought that this facial expression would conflict with the goal of this experiment in which the constant emotion of a vocal expression is expected to lead to varying perceived emotions, depending on the auditory context. Another method is the affective grid (Russell, Weiss, Mendelsohn, 1989), but during the pilot test it was reported by almost everyone, about five different persons participated in the pilot tests, that the size of the seven-by-seven matrix was really bothering them, looked overly complex, so much that it annoyed them. Since in web-based experimentation quitting the test is just one click away, bothering users is not the wisest thing to do. Since none of these methods seemed satisfactory to our needs, the same, but simple, measure was reconsidered.

Gliem and Gliem (2003) point out that single item measures do not allow for assessing the internal validity of the measure, although one could use this argument against the single item measures per variable discussed in the previous paragraph as well. Care should be taken, therefore, in interpreting the results as correct values of actual valence and arousal. To test the hypotheses as stated in this thesis, however, detecting change in response to an emotion rating is already sufficient. It was thus decided to go for the simplest of all measures: one scale per measure. Valence using a 7-point bipolar scale with labels 'Negatief' (negative) and 'Positief' (positive) and arousal using a similar scale, but with the labels 'Passief' (passive) and 'Actief' (active). Additionally, synonyms for each label were provided, as in Experiment 2.

Participants

In total, data from 25 persons was analysed. Most of them, 14, being 21-30 years old, one younger, and 10 older. Eleven females and 14 males participated (median: 21-30). Education levels varied rather much between one participant with vocational training level (VMBO), and two participants having finished post university studies. Most, eight participants, however, had an HBO (Bachelor) degree (translates literally to: “higher professional education”).

Not everyone was able to rate all combinations of utterance x context, only 14 participants rated all samples. From 11 participants, however, we were able to obtain emotion x contexts values, since average values were used to analyse the variance in emotion x context analysis. Therefore not all utterances had to be rated. From seven participants, however, too few data was obtained for a within-person test.

Participants were approached via different mailing lists and newsgroups, including the internal Philips Research newsgroup, the mailing list for students at the Human-Technology Interaction department at the Eindhoven University of Technology, and a mailing list of first year students in Communication Sciences at Tilburg University.

Apparatus

An on-line questionnaire was prepared, comparable to that used in the previous experiment. The utterances used were recordings made during the first experiment and encoded as 128 kb/s MP3 at 44.1 kHz, 16 bit. The on-line questionnaire was programmed in

PHP, and audio playback relied on a small Flash-based player[9]. All randomization was done by the built-in PHP 4 random function.

Aside from a silent context (or no context), there were four auditory contexts created for this experiment. Two intensity levels of white noise, 71.7dB and 79.1dB, were used and two emotional contexts, one positive, 62.3dB, and one negative, 67.8dB[10].

Both emotional sounds were composed of various stock materials and television samples. The positive context included mainly laughter, whereas the negative sound included mainly crying sounds[11].

Procedure

Participants were able to participate wherever they wanted, as the test was internet based. When opening the provided URL in their favourite browser, a short introduction was given. The goal of the experiment was clearly stated, namely investigating the influence of background sound on the emotional perception of utterances. Participants who wanted to proceed could report sex, age, education, and whether they were listening via headphones or not. After registering, a sound check was started, giving participants also the chance to adjust their audio-equipment's intensity to the volume of the samples.

After the equipment test, the actual experiment started. The experiment was divided in ten groups of ten sentences each. Per group of ten sentences one environmental sound was presented. Sentences in the groups were randomized, but it was ensured that in none of these groups two utterances were produced by the same speaker. This to ensure that participants were unable to compare the neutral voice with the emotional voice of that same person directly. Besides, the order was randomized for both the environmental sounds and the order of the sentences within the groups.

Each utterance group started with a rating of the environmental sound, which was repeated continuously in the background, using eight bipolar scales, comparable with those used in Experiment 1. After rating the environment, the ratings of the separate utterances started. Each utterance was repeated three times, after which the user had to rate it on both a valence scale and an arousal scale. If desired, the user was given the option to repeat the

9 XSPF Web Music Player (Flash) - http://musicplayer.sourceforge.net/
10 Intensity values as reported by Praat, based on analysis of the 128kb/s compressed MP3 files
11 All sound are available from http://www.murb.nl/projects/2008/speechemo/

sample. After completion of all the utterances within that group, the environmental context sound was stopped. This procedure was repeated for each of the 10 groups of sentences, after which the participant was thanked for participation. It took most participants about 25-30 minutes.

4.3 Results

The raw results contained ratings of contexts on eight semantic differential scales, and ratings of 16 utterances in different contexts on valence and arousal scales, which were the main dependent variables in this experiment. First, the results of the auditory contexts will be reported, followed by an analysis of how these auditory contexts affected the three types of utterances: negative, neutral and positive. The Lombard utterances will be discussed in Section 5. Although internally all semantic differentials were recorded as values between -3 and 3, values were converted to scales ranging from -1 to +1 by dividing these values (or averaged values in case of the context ratings where emotion was assessed using multiple scales), by 3.

Analysis of the contexts

As in previous experiments, emotion was modelled using the two dimensional valence-arousal model. To measure valence and arousal values of the environmental contexts, multiple bipolar scales were used, like those used in Experiment 1. To verify the internal consistency of both measures, the degree to which the separate questions that were expected to indicate the same emotional component co-varied, the Cronbach's alpha was calculated. The Cronbach's alphas of both the valence scales (α = . 89, n = 6), and arousal scales (α = . 86, n = 4) were found to be good. Hence, only the scale's values for valence and arousal will be used in the analysis that will follow. For further details, please consult Appendix C. Now, the differences in the contexts will be analysed.

Regarding the valence component in the context ratings, Mauchly's test indicated that the assumption of sphericity, i.e. that all levels compared have comparable variances, was violated, χ^2 (9)=23,247, p < .01. To correct this, the Greenhouse-Geisser estimates of sphericity were used (ε = .75). The results show that the level of valence varied significantly with context, F (2.72, 65.21)=81,510, p < .001, r = .87.

For the arousal component in the context ratings, sphericity was safe to assume (p >

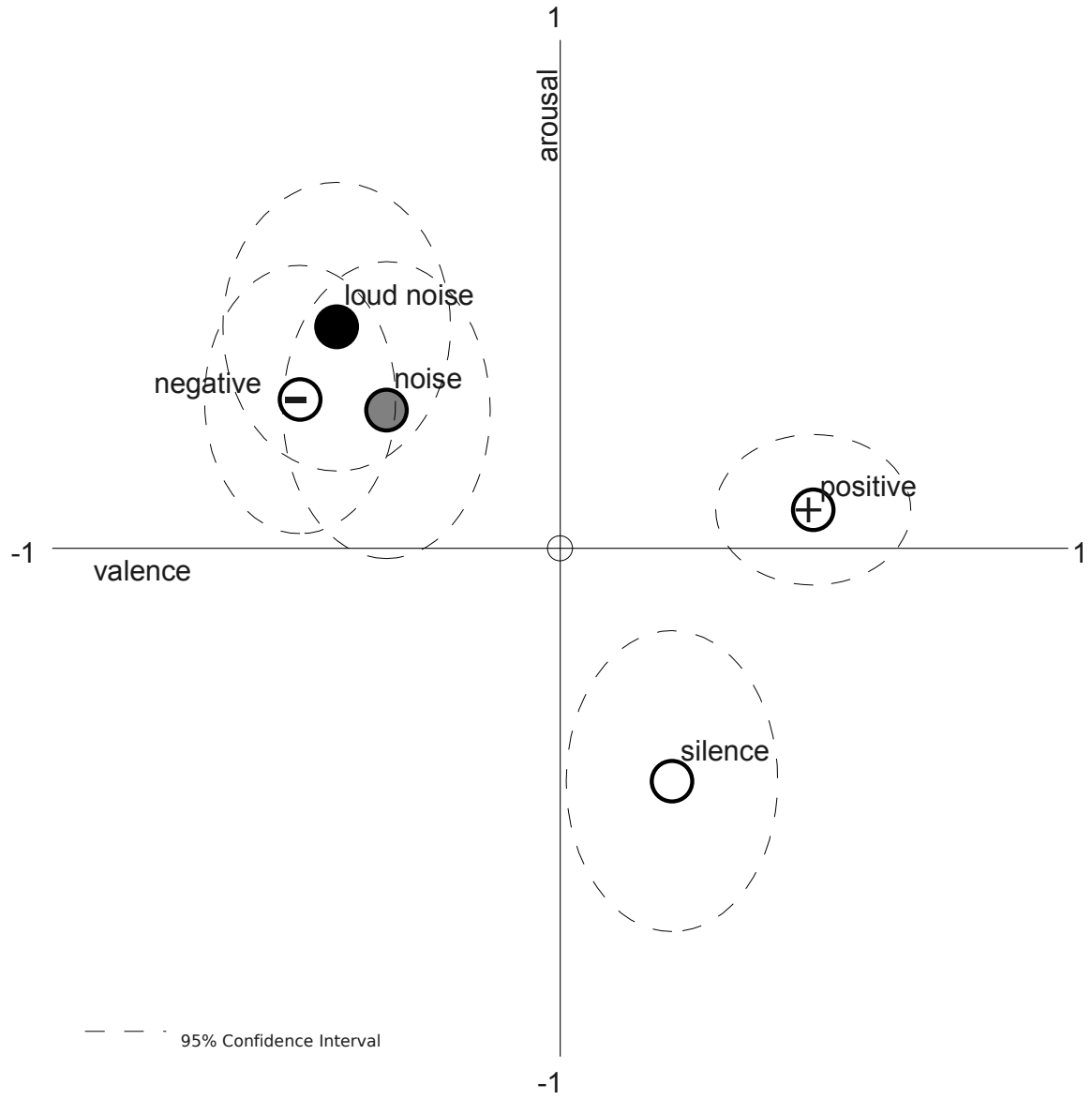

Figure 9 . The position of the five different contexts within the two dimensional emotion space as obtained in Experiment 3.

0.05) . The results show that also the reported level of arousal varied significantly with context, F (4, 96)=33.97, $p < .001$, $r = .75$.

Figure 9 gives an overview of the position of the five context types in the two dimensional emotion space. The circles around the centre dots denote the 95% confidence intervals[12]. It can be seen that the two noise contexts and the negative context are clustered together, and that their 95% confidence intervals overlap; the emotional perception of these three contexts was thus not significantly different.

The negative and two noise contexts were positioned within the negative-aroused

12 Confidence intervals are based on the standard error on the valence and arousal dimensions. The circle shape is an interpolation, assuming that emotions in the two dimensional model roughly follow a circle.

quadrant. Descriptive labels for their position in Russell's circumplex model of affect are 'angry', 'distressed' (Russell, 1980). The position of the positive context is best described with the label 'happy' within Russell's circumplex model of affect. And last, the silence sound approaches the label 'calm'.

Analysis of the emotional utterance judgements when influenced by context

Three categories of utterances have been studied: utterances which express positive emotions, utterances which express negative emotions and last, neutral utterances that express no emotion. In total, participants were able to rate 16 different sentences (four negative, four positive, and eight neutral). The ratings of the different sentences within the same emotion category were averaged for each person within each context. The reason to do so was that with only 25 participants the power of the test, i.e., the chance to detect a genuine effect, would be too low when the effect of context on each sentence was studied separately. More important, subject of study was the effect of context on emotional utterances, not the effect on specific utterances. It should be noted, though, that the internal consistency was not very high for the emotional utterances as indicators for each emotion; both on the valence and arousal dimensions for both the negative and positive conditions, the calculated Cronbach's alpha varied between $.60 < \alpha < .70$. The internal consistency of the neutral sentences was considerably higher for both valence and arousal, $\alpha > .80$.

The sphericity assumption for valence could not be met, neither for the variation between contexts ($\chi^2(9)=34.75$, $p < .001$), nor for the variation between emotion categories ($\chi^2(2)=22.58$, $p < .001$). To correct this, the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity ($\varepsilon = .63$ for the contexts and $\varepsilon = .62$ for emotions). The sphericity assumption could neither be met for the arousal values, nor for the context factor ($\chi^2(9)=46.80$, $p < .001$), nor for the emotion factor ($\chi^2(2)=23.38$, $p < .001$). Again the Greenhouse-Geisser estimates were used to correct for this ($\varepsilon = .48$ and $\varepsilon = .61$, respectively).

The results show a main effect of the uttered emotion. The original emotion significantly influenced the perceived emotion, both on the valence and the arousal component, $F(1.23, 29.53)=40.13$, $p < .001$, $r = .78$ for valence and $F(7.49, 23.26)=80.71$, $p < .001$, $r = .87$ for arousal.

A main effect of context on the judged emotion was also found, again both when

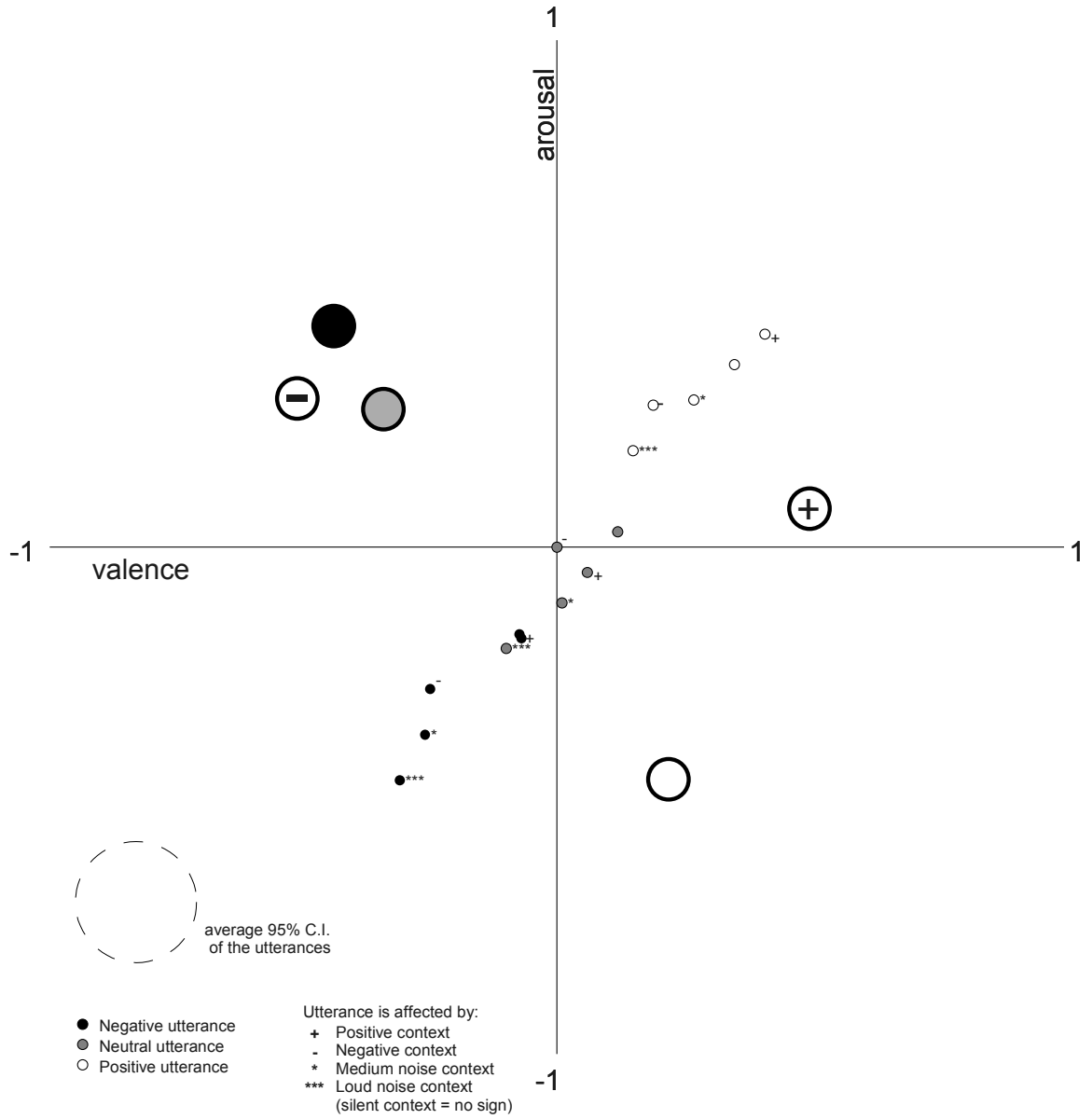

Figure 10 . Position of the utterances as influenced by the context. Context is denoted by the larger circles, as in Figure 9. The smaller circles denote the utterances as influenced by the contexts. The colour indicates the original emotion of the utterance, whereas the symbols denote the context it is affected by.

considering valence, F (2.50, 60.9)=6.177, p = .002, r = .41, and arousal, F (1.94, 46.44)=8.79, p = .001, r = .49. The effect of context on the perceived emotion was thus substantive. Last, the interaction effect, utterance x context, was not significant (p > .05).

Figure 10 shows in more detail the perceived emotions of the different utterances in different environmental contexts. It is interesting to see that the several levels of noise seem to shift the ratings to a lower perceived level of arousal (a shift that is largest when the noise is loudest), thereby "pushing" the target utterance away on the arousal dimension.

On the other hand, emotional environments seem to have an attracting effect on the perceived valence component of emotion; the emotional sentences played in a positive

context were perceived as more positive, and similarly, sentences played in a negative context are perceived more negative. This effect seems to be smaller, however, for neutral sentences. Note that the differences between the perceived emotions are either insignificant ($p > .05$) on the valence component only (silence-positive), the arousal component only (silent-negative) or both components (the other combinations).

4.4 Discussion

This experiment was set up to test the first hypothesis, stating our expectation that different environmental contexts influence the perceived emotion. Based on the results of the ANOVA we have found evidence supporting this hypothesis.

Although evidence was found in support of the hypothesis that the auditory environments influence the perceived emotion, one can see that, based on Figure 10, the shift of the perceived emotion is not in line with the predicted change based on the intuitive guess and the work of Russell and Fehr (1987). Instead of a 180 degrees shift away from the environment's (or anchor) emotion, it is more of a 90 degrees change; changing more towards the valence value of the environment, but moving away along the arousal dimension.

It should be emphasized that this is not a falsification of the work of Russell and Fehr (1987). Despite the similarities, there is also a major difference, which is expected to be the reason why the results cannot be compared. In Russell and Fehr, an earlier rated facial expression influenced a comparable second facial expression. In this experiment, on the other hand, different contexts were used to influence the perception of a vocal expression which cannot be compared directly. In addition, Russel and Fehr's modality was others than the one used in our experiments, visual stimuli versus auditory stimuli. This change in modality, however, is not expected to be the main difference causing the change in results.

It is thought that the results of Russel & Fehr (1987) are due to a temporal order effect, altering the person's emotional reference comparable with experiences that something may look black when compared to white, but turns out to be grey when compared to 'real black', whereas the influence of context is of a different nature. Considering this observation, also some of the variation in the data obtained in the experiment described in this thesis may be attributed to similar order effects as found by Russel & Fehr. Remember that utterances in the experiment described here, were presented in some sequential order to the same participant. An utterance played before, may thus have influenced how the utterance was

perceived and thus rated. Because the order of presentation of the stimuli in this experiment were all randomized, however, these order effects are not expected to have influenced the outcomes obtained in our experiment.

The main reason for not adhering to the same type of set up as Russell and Fehr are practical in nature. First of all, sound is a lot more difficult to control than visual stimuli. It requires rooms with low noise levels. The solution used by Russell and Fehr using, e.g., open days of the faculty to gather participants, was therefore not possible. To solve this, the internet as platform was chosen. While this may not have guaranteed a low noise level at all places, the types of noise, if there, were at least more random.

Although the relevance may not be fully warranted, given the previous observation about the incompatibility of the results, the obtained results put the comments of Ekman and O'Sullivan (1988) on the work of Russell and Fehr (1987) in an interesting new light. Their argument was based mainly on their assumption that only emotionally weak expressions would be susceptible to perceptual shifts. Neutral utterances in the experiment described in this thesis were, however, less affected by the different environmental contexts, when compared to the utterances featuring emotions. This cannot be explained by the data obtained, but maybe emotional speech is more ambiguous than neutral speech?

In the next experiment, the main focus will be on Lombard utterances, which are expected to vary mainly on the arousal axis in the two dimensional valence-arousal model. It is hoped that the results of this experiment help in understanding the effect of the different types of context sounds better.

5. Experiment 4: Lombard speech and Context

5.1 Introduction

In Experiment 3 the effect of background sound on the emotional perception of emotional speech was investigated. The main purpose of the experiment described here is to verify the second hypothesis: speech recorded in noisy environments, so-called Lombard speech[13], is perceived less neutral outside the noisy context; i.e. more negative and more aroused. The two levels of noise will be used to create a medium aroused, and highly aroused auditory environment. It is expected, based on the literature reviewed in the introduction of the previous experiment and the results obtained in the previous experiment regarding the arousal component, that the high-noise (and thus highly aroused) condition will lower the perceived arousal of the speech sample, whereas the low noise, and especially the no-noise background presentation will lead to increased perceived arousal.

5.2 Method

Design

The design used in this experiment was a full factorial design in which both environmental sounds and utterances were varied as the independent variables. The environmental sound conditions were three different intensity levels of noise: no sound, soft noise and loud noise. Four utterances, see for the method of recording the apparatus section at page 40, were used, two of which were Lombard speech, two of which were neutral.

The dependent variable was the rated emotion of the utterance played in context and the rated mood of the environment, as in Experiment 3.

Participants

The initial group of participants was the same as in Experiment 3, as the stimuli were presented in a joint test. Since less data were required for this analysis, data of 26, in contrast to 25 of Experiment 3, were suitable for a within-person test. Most of them, 15 participants, being 21-30 years old (median group 21-30). 1 was younger (11-20), 10 participants were older (30+). Eleven females and 15 males participated. Education levels varied considerably

13 See for more information “Box 4: Stress and the Lombard Effect” at page 8.

between at lowest vocational training levels (Dutch: "MBO"), and at highest, post university graduates. 9 participants, however, had finished HBO-level education (Bachelor level, translates literally to: "higher professional education").

Apparatus

Generation of the stimuli. There are several stressors that can cause the Lombard effect (see box 4 at page 8). It is often assumed that the Lombard effect is mainly based on the tendency of people to talk louder when they are not well understood (see Junqua, Fincke & Field, 1999). Other than in Junqua et al. (1999), the communication system was set up between two real persons (like Bořil, Bořil and Pollák, 2006). The main reason to do so was practical in nature: there was no automatic voice controlled menu available. The communication server and client were created using PureData[14], a graphical programming environment based on patches, like Max/MSP, using the netsend~ and netreceive~ objects[15], allowing for sending uncompressed audio data in real time over a TCP/IP based network. The server allowed for controlling the noise level, and at the client side, microphone input was stored as a wave file. Because earphones were used at either side, the microphone captured only speech, and no noise.

The participant was required to take a seat in a separate room and was provided with a headphone. A microphone was positioned in front of the person. In another room, a confederate was seated. The participant was asked, by the confederate, to read out aloud the same seven neutral sentences as those used in Experiment 1. After finishing reading, the noise level was increased, and the participant was required to reread the text. This was repeated for four different levels of noise. If the confederate could not hear the other person properly, the person was asked to repeat the sentence. Only the last two sentences of the list were the target sentences, intended for use in this experiment.

Material was recorded at a empty office within the Philips research department using a AKG CK 31 cardoide microphone mounted to a AKG LM 3 connector. The signal was pre-amplified using a RME QuadMic pream. The two channel signal (XLR) was then wired to a Dell laptop (onboard soundcard). The audio was encoded as 48 kHz, 16 bits, uncompressed audio. It was only possible to record in mono, so no phase inversion could take place which might have reduced some of the unwanted noise. On the other hand, it was ensured that the

14 PureData - http://puredata.info/
15 netsend~ for Max/MSP and Pure Data - http://www.nullmedium.de/dev/netsend~/

input signal was at an optimal level[16]; the noise recorded due to electronic circuits, etc., is negligible. The microphone, however, did record some of the noise presented to the participants in the noisier stages. The audio was cut in separate parts using Audacity 1.2.6.

Since Lombard speech may be considered 'screaming', which in turn may be related to something like anger (or joy, see also box 3), the selected speech samples have been analysed and compared with angry and joyful speech (Table 3), which it might resemble, like in Experiment 2.

Table 3.
*Change in speech prosody. Comparing in the left column non-Lombard speech and Lombard speech obtained for this experiment and in the right column predicted changes based on a literature review of Scherer (2003) for speech expressing anger and joy/elation. ** = p < .05, * = p < .10.*

	Lombard		
	Present study (n=2)	**Scherer (2003) Anger**	**Scherer (2003) Joy/ elation**
Intensity (dB)	+15.18 **	↗	↗
F0 mean (semitones)	+7.27	↗	↗
F0 stdev (semitones)	+1.34	↗	↗
F0 range (semitones)	+4.60	↗	↗
Contour trend (semitones/s)	-1.75 *	↘	
Relative high frequency energy	+0.21	↗	(↗)
Speech/Articulation Rate (syl/s)	-0.24	↗	(↗)

Equipment & procedure used in the actual experiment. Equal to that of Experiment 3.

5.3 Results

For this experiment a subset of the auditory environment sounds used in the previous experiment was used; only the silence, medium noise and loud noise stimuli were used. For more information on these contexts, consult the results section of Experiment 3.

16 One sample was recorded with a level that was somewhat higher than optimal, resulting in clipping.

The assumption of sphericity was met for all results being analysed in this section, based on Mauchly's test ($p > .05$). The main effects of the different noise stimuli on the valence component of the perceived emotion was insignificant, F (2, 50)=.568, *ns* . Neither was the effect of the actual utterance on the perceived valence, F (1,25)=1.618, *ns* .

The arousal component did vary significantly with both context and utterance-type. The results show that the perceived arousal of the Lombard sentences was significantly different from the non-Lombard sentences, F (1,25)=152,781, $p < 0.001$, $r = .93$. Additionally, the acoustic context also significantly influenced the perceived arousal, F (2,50)=9.676, $p < 0.001$, $r = .51$. This indicates that when the type of sentence is ignored, the perceived arousal varies with the context in which it is presented. Post hoc t -tests reveal that the main difference is between the noise and no-noise contexts; the differences between the soft and loud noise are insignificant ($p > .05$), whereas the difference between soft noise and silence, and loud noise and silence are significant ($p < .05$ and $p < .01$, respectively).

The utterance type x environment interaction was only significant for the perceived arousal component, F (2,50)=3.85, $p = .03$, $r = .32$, indicating that the effect of the environment on arousal differed between the two sentence types.

Figure 11 describes how the two types of speech (normal and Lombard) are affected by the three intensities of noise. It is clear from the 95% confidence intervals that the Lombard speech is hardly affected by the type of background stimulus. Within Russell's (1980) circumplex model of affect, this position is best described with the labels 'alarmed' and 'tense'.

Neutral, or non-Lombard, speech, however, is significantly different, as found also in the post hoc tests; neutral sound in a silent context is perceived as significantly more neutral (relatively more aroused) when compared to noisy contexts. With increasing intensity, the perceived emotion seems to move into the direction of what is labelled 'droopy' and 'bored' in the circumplex model of affect (Russell, 1980).

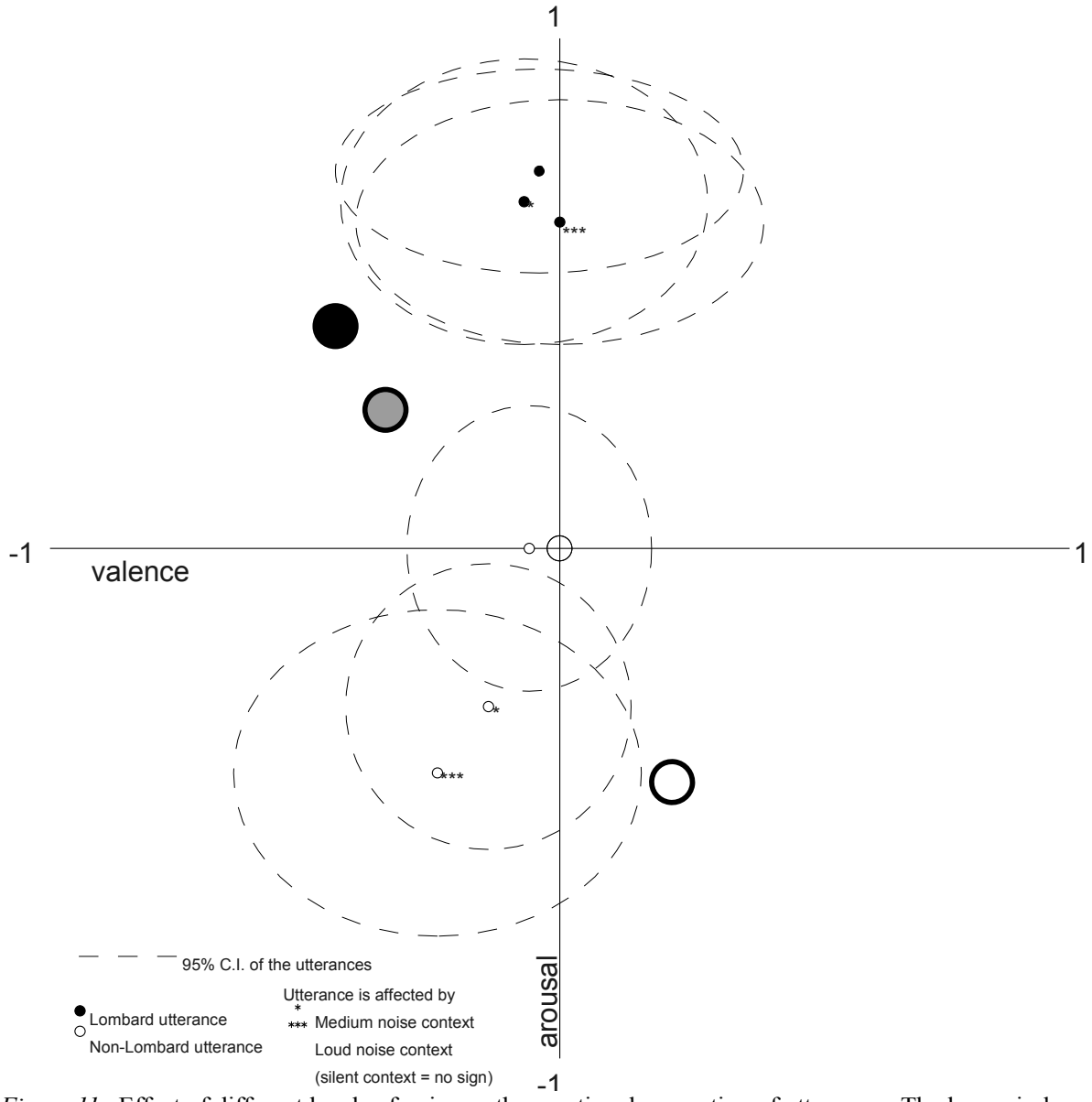

Figure 11 . Effect of different levels of noise on the emotional perception of utterances. The large circles denote the perceived emotion of the environments, whereas the smaller circles denote the utterances influenced by either a medium noise context, a loud noise context or silence.

5.4 Discussion

This experiment was set up to test the second hypothesis, stating our expectation that speech recorded in noisy environments will be perceived as less neutral when listened to outside the noisy context. We expected that this type of speech would be rated as more aroused in a silent context. Although the results of the ANOVA confirm that the influence of the environment was significant, there was also a significant interaction effect. The actual effect is clear when looking at Figure 11; the Lombard speech is hardly affected by type of auditory context in which it is presented, hence the hypothesis is rejected.

Maybe Ekman and O'Sullivan's (1988) argument against the work of Russell and Fehr (1987) is applicable here. The recorded Lombard sentences were very much activated, and the fact that the two persons were 'screaming' was quite clear. On the other hand, also the loudest noise context was very much activated; i.e. very loud. Some participants even complained in the feedback form that they were almost unable to hear voices, and that the noise levels almost hurt.[17] It might be interesting to repeat this experiment with various levels of Lombard speech.

The main reason to investigate the Lombard effect was that the highly aroused speech of a screaming person in a, to the perceiver, silent environment was expected to lead to an awkward percept. It was expected that this could be demonstrated by comparing the emotional perception of Lombard speech in different contexts, but that was not the case. This does, however, not rule out that Lombard speech in environments with noise reduction applied does lead to confusion in the communication.

There are several reasons why the anticipated effect did not occur. First of all, there is a possibility that a ceiling effect occurred in measuring the arousal. Note that the arousal values obtained for Lombard speech are the most extreme values measured in this experiment. In contrast to the other samples, the aroused Lombard speech might have been experienced as highly aroused, no matter in what context it was presented. A second alternative explanation may be that the neglected dimension of dominance would make a difference in interpretation, a dimension that is incorporated in some models of emotion. Last, it could simply be that the perceived arousal is not influenced by any form of context.

The significant shift in perceived arousal for the non-Lombard speech, however, is still interesting. Although this finding may have little relation to the original motivation of this research, that of environmental noise suppression, it does underline the importance of good quality sound in transmission of speech. The added noise decreased the perceived activation, resulting in a somewhat dull perception of the speaker.

As a last note, one may ask whether people who know that the environmental noise is suppressed at the receiver's side by technical means still utter typical Lombard speech. Lombard speech is thought to be caused by two types of stressors, a second order stressor and a third order stressor (Murray, Baber & South, 1996; see also Box 4). The second order

17 Participants were informed about the type of sounds they would hear. In case of too loud noise they were also free to adjust their volume appropriately.

stressor relates to the stress experienced by a speaker because of the noise in the environment. The third order stressor relates to the changes in speech that are more cognitive in nature, e.g., speaking louder in reply to a receiver's request to 'speak up'. Both stressors were present when the stimuli used in this experiment were generated; noise was presented as a second order stressor, and the confederate would[18] ask the other person to speak up when he could not hear him or her, acting as a third order stressor. The influence of the third order stressor, however, maybe reduced in a communication system with noise-suppression capabilities, if the sender is aware of this functionality of the system. In that case, the receiving end will probably not complain about poor reception quality, and the sender may be less uncertain about how he or she is being received when he or she knows that environmental noise is suppressed. It is thus no longer required to speak up loudly, i.e., to utter Lombard speech, for reasons of speech intelligibility at the receiver's side. It remains yet to be seen, however, whether the majority of the users actually understands that their signal to noise ratio is much improved at the receiver's side, and act accordingly; i.e., start talking at a more normal volume, even in noisy conditions. Additionally, the second order effect, the stress caused by the noise in the room, will also still be present, affecting how the sender speaks. It is, therefore, expected that the results are still reasonably valid, even when the future application of environmental noise reduction is considered.

18 The voice of the confederate has not been recorded, nor has the confederate kept track of the times he had to ask a participant to speak up.

6. General discussion

The main topic of this thesis was the role of the auditory environment in mediated vocal communication. It was hypothesised that although emotional information is contained in speech, the perceived emotion is affected by the environment.

To investigate this hypothesis, the commotion model was used (Scherer, 1998; Scherer and Zentner, 2001). In this model two main influences of the environment were discerned: that of the environment affecting the interpretation of a speaker's emotion and that of the environment affecting the speaker, resulting in a change in the produced and, hence, perceived voice prosody (or 'symptoms'). Both routes were studied in two separate experiments, Experiment 3 and 4. The experiments completed before experiment 3 were necessary to generate the experimental material. Although especially the method of generation was interesting, it is not the main subject of this thesis, and will therefore not be discussed in great detail in this general discussion.

The main finding of this study was that the perceived arousal decreased with increased arousal of the context, and vice versa. Perceived valence, however, increased when the valence of the context was higher, and decreased when the valence of the context was lower. As discussed in the discussion sub-section of Experiment 3, the latter finding was not as expected based on the study by Russell & Fehr (1987), nor from the more intuitive prediction.

There are several explanations for the results obtained. First of all, audio might simply behave differently than visual perception of emotion. This seems unlikely, but a closer replication (note that this was not the goal of this research) of Russell and Fehr's (1987) in the auditory domain could be helpful in improving our understanding.

A second explanation may be that listeners somehow 'average' the valence of the context with the valence of the environment. Maybe context triggers some process that emphasises the prosodic features in speech (remember that there is ambiguity in emotion). Or, more like music in a film, it has the function to explain the emotional state of a person. These explanations, however, would leave the shifts on the arousal axis, which lead to a shift away from the context's arousal of Experiments 3 and 4, unexplained.

Possibly, the various levels of noise were not experienced as an emotional environments, or follow somehow different 'rules'. Maybe arousal in general follows different rules than valence. Barrett, Mesquita, Ochsner and Gross (2007) already emphasized that it is debatable whether arousal is part of one's core affect (see also Box 1). But it is hard to draw any conclusions on why, after only two experiments. Also the results of Russel and Fehr (1987) do not seem to confirm this idea.

It should be noted that the original interpretation of context may have been too simplistic. Therefore, the results of Russel and Fehr (1987) and the results of this research are less comparable than originally thought, because different things were varied in the two setups. In Russell and Fehr, two faces were placed next to each other and participants were instructed to first rate the left face, then followed by the right face. Although it was thought that the anchor face, the left face, could be considered more or less like a bystander in the environment of the target face, there might be an alternative explanation: a temporal order effect; the earlier presented stimulus influenced the stimulus presented directly afterwards. The earlier perceived stimulus is used as reference to compare the second stimulus against.

As said, it was thought that an auditory environment could work in a similar way as the anchor faces of Russel and Fehr (1987); the environment is always there and can therefore, in general, be perceived before the speaker is perceived. Also in the experiments described in this thesis, the participant is asked to rate the environment first, and only then, while having the environment available as background sound, to rate several utterances. The main difference is that not only one utterance is judged after having judged the environment, but multiple utterances uttered by multiple persons in the same auditory environment. The environment may have become, especially after a few utterances being rated, more of a background sound than a stimulus to compare the target utterance against. Nonetheless, both a shift by an earlier presented stimulus, and a shift by background sound are context effects, that matter in mediated communication.

Generalizing these observations, it seems that in everyday communication the interpretation of and observed emotion, commotion, depends on more than just induction, empathy and contagion as recognized by Scherer and Zentner (2001). Stimuli that are perceived before, and that are independent from the sent emotion, may also affect how follow up, temporarily ordered, experiences are perceived emotionally. A news item, for example, about a child that died, is expected to affect how a following unrelated report of a soldier that

died is perceived emotionally. Russell and Fehr (1987) have shown that a happy face displayed just before a neutral face influences the emotional precept of the neutral face. Today's media allow for instantaneous changes in scenes and context presented to the observer. Also in communication, users may switch quickly between calls. Changes in what is presented before the target item, is also expected to influence the emotional percept of the target item. Further analysis of these issues is, however, beyond the topic of this thesis.

One aspect that has been left unmentioned, is the generation of the stimuli. Although it does not threaten the claims made about the relativity of the emotions, it should be noted that emotional utterances were recorded in a silent office where no background sound was provided to the speaker when recording the stimuli. It is somewhat important, however, when it is external validity that is concerned, as the experiment was set out to test the effects of background suppression perceivable on the receiver's end only. Maybe negative stimuli should have been recorded with crying sounds in the background, and similarly, positive speech stimuli should have been recorded with laughing people in the background, making the emotional percepts also more stable. It is thought, however, that the effect of this is negligible.

7. Conclusion

In this thesis the influence of the auditory environment on the emotional perception of speech in mediated communication was addressed. The motivation of this study was the development of techniques that enable suppression of environmental sound, communicating only the sound coming from the speaker.

There is still much uncertainty on how the process of emotional perception works, but regarding the emotional perception in the two dimensional valence-arousal model there does not seem to be much reason for not applying the more advanced environmental noise suppression algorithms when these algorithms lead to an improved intelligibility. Although shifts have been perceived, the shifts are not huge shifts from a negative percept to a more positive percept or the other way around. A positive person speaking, for example, calling from a cheerful environment, is perceived by the receiver not hearing the cheerful context as only a little less positive than he or she would be experienced within the cheerful context. A person talking in a very noisy environment, but heard in isolation, is not heard that different either, although it maybe worthwhile to study the effects of context on speech recorded in less noisy conditions. So far, no evidence was found that could support a claim that removing the environmental sound would lead to impoverished communication. On the other hand, only a small aspect, that of emotional perception, has been addressed here. Additionally, there are some methodological issues that may not have allowed for detection of a change.

The findings of this study left interesting food for thought: Would it be possible to replicate Russell & Fehr's results with auditory stimuli only? What about the multi-modal effects, when auditory stimuli are combined with visual stimuli (e.g. auditory context, visual subject, or audio-visual subject)? Maybe perceived arousal of aroused speech in noisy contexts was rated relatively stable because of ceiling effects? And how does order of presentation affect how we perceive stimuli emotionally? Although the combinations of environment and utterances tested were, with hindsight, not fully representative for expected future situations that may arise when environmental noise is suppressed, the results do indicate that environmental sound does play a non-negligible role in how we perceive persons emotionally. Maybe the context is the message[19]?

19 Free after McLuhan's “the medium is the message” (McLuhan, 1964)

References

Baber, C. & Noyes, J. (1996). Automatic speech recognition in adverse environments. *Human Factors* , *38* , pp. 142-155.

Barrett, L. F. (2006). Solving the emotion paradox: categorization and the experience of emotion. *Personality and Social Psychology Review* , *10* , pp. 20-46.

Barrett, L. F., Mesquita, B., Ochsner, K. N. & Gross, J.J. (2007). The experience of emotion. *Annual Review of Psychology* , *58* , pp. 373-403.

Boehner, K., Depaula, R., Dourish, P. & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies* , *65* , pp. 275-291.

Bořil, H., Bořil, T. & Pollák, P. (2006). *Methodology of Lombard Speech Database Acquisition: Experiences with CLSD* . Retrieved from http://noel.feld.cvut.cz/speechlab/publications/045_lrec06.pdf.

Boves, L. W. (1984). The phonetic basis of perceptual ratings of running speech (Doctoral Thesis). Dordracht - Holland: Foris Publications.

Bradley, M. M. & Lang, P.J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry* , *25* , pp. 49-59.

Brown, P. & Fraser, C. (1979). Speech as a marker of situation. In Scherer, Klaus R. and Giles, Howard (Ed.), *Social markers in speech* . Cambridge: Cambridge University Press. pp. 33-62.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. & Weiss, B. (2005). A database of German emotional speech. In *INTERSPEECH-2005* (pp. 1517-1520). Lisbon: ISCA.

Cauldwell, R. T. (2000). Where did the anger go? The role of context in interpreting emotion in speech. In *SpeechEmotion-2000* (pp. 127-131). Newcastle, Northern Ireland, UK: ISCA.

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

Desmet, P. M. A. (2003). Measuring emotion; development and application of an instrument to measure emotional responses to products. In Blythe, M. A. and Monk, A. F. and Overbeeke, K. and Wright, P. C. (Ed.), *Funology: from usability to enjoyment* . Dordrecht: Kluwer Academic Publishers. pp. 111-123.

Desmet, P. M. A., Hekkert, P. & Jacobs, J.J. (2000). When a car makes you smile: Development and application of an instrument to measure product emotions. *Advances in Consumer Research* , *27* , pp. 111-117.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication* , *40* , pp. 33-60.

Ekman, P. & O'Sullivan, M. (1988). The role of context in interpreting facial expression: comment on Russell and Fehr (1987). *Journal of Experimental Psychology. General* , *117* , pp. 86-98.

de Gelder, B. & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion* , *14* , pp. 289-311.

Gliem, J. A. & Gliem, R.R. (2003). *Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales* . Retrieved from http://alumni-osu.org/midwest/midwest%20papers/Gliem%20&%20Gliem--Done.pdf.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W.C. (1995). Multivariate data analysis. Upper Saddle River, NJ, USA: Prentice-Hall.

Johnstone, T. & Scherer, K.R. (2000). Vocal communiciation of emotion. In Lewis, M. and Haviland-Jones, J. (Ed.), *Handbook of Emotions* . New York: Guilford Press. pp. 220-235.

Junqua, J. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication* , *20* , pp. 13-22.

Junqua, J. C., Fincke, S. & Field, K. (1999). The Lombard effect: a reflex to better communicate with others in noise. In *ICASSP '99 Proceedings* (pp. 2083-2086, vol. 4). Phoenix, Arizona: IEEE.

Kenealy, P. M. (1986). The velten mood induction procedure: A methodological review. *Motivation and Emotion* , *10* , pp. 315-335.

Krahmer, E. & Swerts, M. (2008). *Displayed, but not felt - production and perception congruent and incongruent emotional speech* . Unpublished Paper.

Krahmer, E., Dorst, J. & van Ummelen, N. (2004). Mood, persuasion and information presentation. *Information Design Journal* , *12* , pp. 219-232.

Laukka, P. (2004). *Vocal expression of emotion: Discrete-emotions and dimensional accounts* . Doctoral dissertation, Acta Universitatis Upsaliensis, Uppsala. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-4666.

Massaro, D. W. & Egan, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review* , *3* , pp. 215-221.

McLuhan, M. (1964). Understanding Media (critical edition by T. Gordon; Dutch Translation 2002). Amsterdam: Gingko Press.

Mozziconacci, S. J. (2001). Modelling Emotion and Attitude in Speech by Means of Perceptually Based Parameter Values. *User Modelling and User-Adapted Interaction* , *11* , pp. 297-326.

Murray, I. R., Baber, C. & South, A. (1996). Towards a definition and working model of stress and its effects on speech. *Speech Communication* , *20* , pp. 3-12.

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

Rosenberg, E. L. (1998). Levels of Analysis and the Organization of Affect. *Review of General Psychology* , *2* , pp. 247-270.

Rottenberg, J., Ray, R. D. & Gross, J.J. (2007). Emotion elicitation using films. In J. A. Coan & J. J. B. Allen (Eds.), *The handbook of emotion elicitation and assessment* . London: Oxford University Press. pp. 9-28.

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

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review* , *110* , pp. 145-172.

Russell, J.A. & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions.

Journal of Experimental Psychology General , *116* , pp. 223-237.

Russell, J. A., Weiss, A. & Mendelsohn, G.A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology* , *57* , pp. 493-502.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice* , *9* , pp. 235-248.

Scherer, K. R. (1998). Emotionsprozesse im Medienkontext: Forschungsillustrationen und Zukunftsperspektiven. *Medienpsychologie* , *10* , pp. 276-293.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication* , *40* , pp. 227-256.

Scherer, K. R. & Zentner, M.R. (2001). Emotional effects of music: production rules. In Juslin, P. N. and Sloboda, J. A. (Ed.), *Music and Emotion: Theory and Research* . Oxford: Oxford University Press. pp. 361-392.

Smith, E. E., Nolen-Hoeksema, S., Fredrickson, B. & Loftus, G. (2002). Atkinson and Hilgard's Introduction to Psychology. Pacific Grove, CA: Wadsworth Publishing.

Steeneken, H. J. M. & Hansen, J.H.L. (1999). Speech under stress conditions: overview of the effect on speech production and on system performance. In *ICASSP '99: Proceedings of the Acoustics, Speech, and Signal Processing* (pp. 2079-2082). Phoenix, Arizona: IEEE.

van den Stock, J., Righart, R. & De Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion (Washington, D.C.)* , *7* , pp. 487-494.

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy* , *6* , pp. 473-482.

Westermann, R., Spies, K., Stahl, G. & Hesse, F.W. (1996). Relative effectiveness and validity of mood induction procedures: a meta-analysis. *European Journal of Social Psychology* , *26* , pp. 557-580.

Wilting, J. (2005). *Visuele expressies van emoties in congruente en incongruente condities* . Unpublished master's thesis, Tilburg University, Tilburg.

Yik, M. S. M., Russell, J. A. & Barrett, L.F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of personality and social psychology* , *77* , pp. 600-619.

Appendix A: Dutch Velten Sentences

Sentence material for the Velten method was obtained from Wilting (2005), who translated the original sentences of Velten (1968) to Dutch. Since the Velten induction was already preceded with a film induction, the amount of sentences was reduced from 40 to 25 for the emotional sentences and to only 5 for the neutral sentences based on random selection.

Positive Velten Sentences

1. Ik voel me best wel goed vandaag.
2. Deze dag zou wel eens een van mijn betere dagen kunnen zijn.
3. Ik heb energie en zelfvertrouwen in overvloed.
4. Ik voel me opgewekt en vrolijk.
5. Ik denk dat vandaag alles verder heel goed zal gaan.
6. Mijn mening over de meeste zaken is weloverwogen.
7. Ik denk dat er mooie tijden aankomen.
8. Ik weet heel goed dat ik mijn doelen kan bereiken.
9. Ik voel me sterk en vitaal.
10. Niemand kan me stoppen vandaag!
11. Ik voel me verbazingwekkend goed vandaag!
12. Ik voel me creatief en inventief vandaag.
13. Ik voel me super!
14. Ik zie alles van de zonnige kant.
15. Ik voel me erg opgewekt en levendig.
16. Ik zie alles scherp en in een nieuw daglicht.
17. Ik kan me goed concentreren op alles wat ik doe.
18. Ik denk helder en snel.
19. Het leven is zo leuk; het geeft me zoveel voldoening.

20. Alles zal vandaag steeds beter gaan.
21. Ik voel me energiek. Ik wil iets doen!
22. Dit is geweldig; ik voel me echt goed.
23. Dit is zo'n dag waarop ik ervoor ga!
24. Ik zit vol energie.
25. God, wat voel ik me geweldig!

Negative Velten Sentences

1. Ik voel me neerslachtig vandaag.
2. Ik voel me best sloom op het moment.
3. Het lijkt wel alsof iedereen energie heeft, behalve ik.
4. Nensen irriteren me. Waarom laten ze me niet met rust?
5. Ik heb het gevoel dat ik nauwelijks vooruit kom.
6. Van een beetje inspanning word ik al moe.
7. Ik voel me vandaag verschrikkelijk moe en alles kan me gestolen worden.
8. Ik begin me slaperig te voelen. Ik dwaal steeds af.
9. Mijn leven is zo vervelend, elke dag diezelfde sleur is deprimerend.
10. Ik ben geen stuiver waard.
11. Ik voel me belabberd. Mijn gezondheid is niet zoals het zijn moet.
12. Niemand begrijpt me als ik klaag of me ongelukkig voel over mezelf.
13. Ik ben onzeker over mijn toekomst.
14. Ik ben moedeloos en ongelukkig met mezelf. Alles is nu slechter dan toen ik jonger was.
15. Zoals ik me nu voel, ziet de toekomst er saai en hopeloos uit.
16. Ik vind het ontzettend moeilijk om belangrijke beslissingen te maken.
17. Ik voel me moe en depressief; ik heb geen zin om iets te doen.

18. Vaak maken mensen me erg boos. Ik ben liever alleen.
19. Ik kan niet goed over mijn problemen praten met anderen.
20. Mensen luisteren nooit echt naar me.
21. Soms wou ik dat ik dood was.
22. Ik geef nergens meer om. Het leven is gewoon niet leuk.
23. Ik heb geen zin om iets te doen.
24. Alle tegenslagen in mijn leven achtervolgen me.
25. Ik wil slapen en nooit meer wakker worden.

Neutral Velten Sentences

1. Aan het einde van het boek vind je de bibliografische gegevens
2. De organisatie heeft verschillende dochterondernemingen.
3. Het bedrijfsplan werd niet veranderd.
4. Het grote herenhuis staat te koop.
5. Het mandarijn is de officiële taal van China.

Appendix B: Measure of emotion

In this appendix it will be described how the eight semantic differential scales (as used in Experiment 1, 3 and 4 were converted to a valence and an arousal value.

The semantic differentials presented were based on Yik, Russell, Barrett (2003), and translated to Dutch (and pairs were counterbalanced): Plezierig (Pleased) – Ellendig (Miserable) [PE]; Verontrustend (Nervous) – Ontspannend (Relaxed) [VO]; Prikkelend (Aroused) – Kalm (Quiet) [PK]; Deprimerend (Depressing) – Stimulerend (Stimulating) [DS]; Rustgevend (Still) – Opwindend (Exciting) [RO]; Verdrietig (Unhappy) – Blij (Happy) [VB]; Kalmerend (Calm) – Beangstigend (Tense) [KB]; Enthousiast (Excited) – Somber (Gloomy) [ES].

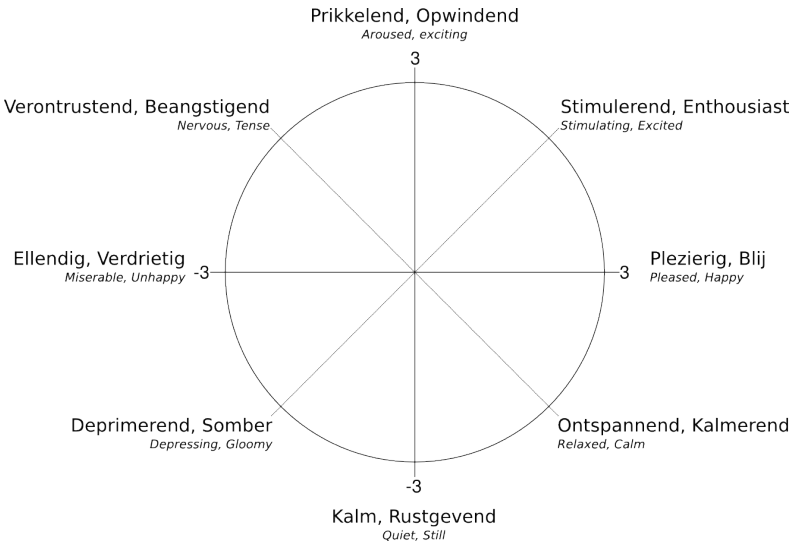

Figure 12: Emotional opposites in the Valence Arousal model.
Adapted from Yik, Russell, Barrett (2003).

Although PE and VB (capitals refer to the first letter of the emotional label in Dutch), corresponded perfectly with the valence axis, and PK and RO with the arousal axis, DS, VO, KB and ES were indicators of both axis to some extent. Since the model is modelled as a circle, the weight assigned to the latter four extremes was $\frac{\sqrt{(2)}}{2}$.

To verify whether the translated bipolar scales were actually in correspondence with each other, which has been assumed so far, the Cronbach's alpha has been calculated for each

measure. For the valence axis this was based on PE, VB, and the weight-adjusted versions of DS, VO, KB and ES. For both Experiment 1, and the shared Experiments 3 & 4 the Cronbach's alpha was found to be .89, which is considered to be a good result (George & Mallerey, 2003, as cited by Gliem & Gliem, 2003). For the arousal axis this was based on the values obtained for PK, RO, and the weight adjusted versions of DS, VO, KB and ES, α =.54 (for Experiment 1) and, α =. 66. These values are considered, respectively, poor, and acceptable. However, by removing the ES and DS values, reliability improved considerably (resp. α =.81 . α = .86) , hence it was decided to exclude these values from the final calculation of the arousal value. The formulas to obtain the valence and arousal values thus became:

$$V = VB - PE + \frac{\sqrt{(2)}}{2} VO + \frac{\sqrt{(2)}}{2} DS - \frac{\sqrt{(2)}}{2} KB - \frac{\sqrt{(2)}}{2} ES$$

and,

$$A = RO - PK + \frac{\sqrt{(2)}}{2} KB - \frac{\sqrt{(2)}}{2} VO \quad .$$

Appendix C: Praat speech analysis script

Note: instead of the f_0-values obtained using this praat script, close copy contours created using GIPOS 2.1 and by an experienced close-copy stylist have been used in analyses reported in this thesis.

```
echo File, f0_mean, f0_stdev, f0_range, f0min, f0max, slope, voicedproportion, intensity,
highfrequencyenergy, length, hammi, pe1000
n = numberOfSelected ("Sound")
for i to n
    sound'i' = selected ("Sound", i)
endfor

for i to n
        select sound'i'
        filename$ = selected$ ("Sound")
        length = Get total duration
        intensity = Get intensity (dB)
        To Pitch... 0.0 50.0 600.0
        f0mean = Get mean... 0.0 0.0 Hertz
        f0stdev = Get standard deviation... 0.0 0.0 Hertz
        f0min = Get minimum... 0.0 0.0 Hertz Parabolic
        f0max = Get maximum... 0.0 0.0 Hertz Parabolic
        f0range = f0max - f0min
        framestotal = Get number of frames
        framesvoiced = Count voiced frames
        voicedp = framesvoiced / framestotal

        #Linear regression on pitch values
        meantime = Get total duration
        meantime = meantime / 2

        s_xx = 0
        s_xy = 0

        for iframe to framestotal
                time = Get time from frame... iframe
                pitch = Get value in frame... iframe Hertz
                if pitch != undefined
                        s_xx = s_xx + ((time-meantime) * (time-meantime))
                        s_xy = s_xy + ((pitch-f0mean) * (time-meantime))
                endif
        endfor
        slope = s_xx / s_xy
        #end of linear regresson on pitch values
        Remove

        select sound'i'
        #relative energy in the highfrequency region
        To Spectrum... Fast
        energyspectrumtotal = Get band energy... 0.0 22050
        energyspectrumhigh = Get band energy... 1000.0 22050
        highfrequencyspectrum = energyspectrumhigh / energyspectrumtotal
        Remove

        select sound'i'
        # hammarberg index
        To Ltas... 100
        ltasmaxa = Get maximum... 0 2000 None
        ltasmaxb = Get maximum... 2000 5000 None
        hammi = ltasmaxa - ltasmaxb
        Remove

        select sound'i'
        # voiced long term average spectrum
        To Ltas (pitch-corrected)... 75 600 5000 100 0.0001 0.02 1.3
        ltaspmaxa = Get mean... 1000 44100 energy
        ltaspmaxb = Get mean... 0 1000 energy
        ltaspmaxa = 10^ltaspmaxa
        ltaspmaxb = 10^ltaspmaxb
        pe1000 = ltaspmaxa / ltaspmaxb
        pe1000 = log10(pe1000)
        Remove

        printline 'filename$', 'f0mean:2', 'f0stdev:2', 'f0range:2', 'f0min:2', 'f0max:2',
'slope:2','voicedp:2', 'intensity:2', 'highfrequencyspectrum:2', 'length:2', 'hammi:2', 'pe1000:2'
endfor
```