Audiovisual prosody in interaction

Barkhuysen, P.

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Audiovisual Prosody in Interaction

Pashiera Barkhuysen

ISBN: 978-90-9023358-1

Print: Ridderprint, Ridderkerk

Cover photo: Mae West Lips Sofa by Salvador Dali

© Orlando Florin Rosu | Dreamstime.com

Cover design: Lennard van de Laar

Audiovisual Prosody in Interaction

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Tilburg,
op gezag van de rector magnificus,
prof. dr. F. A. van der Duyn Schouten,
in het openbaar te verdedigen ten overstaan van een door
het college voor promoties aangewezen commissie
in de aula van de Universiteit

op vrijdag 3 oktober 2008 om 14.15 uur

door Pashiera Nicolet Barkhuysen,

geboren 6 maart 1972 te Dwingeloo

Promotores:

Prof. Dr. M. G. J. Swerts

Prof. Dr. E. J. Krahmer

Promotie-commissie:

Prof. Dr. E. André

Prof. Dr. A. A. Maes

Prof. Dr. C. Gussenhoven

Prof. Dr. B. Granström

Prof. Dr. D. R. Ladd

Dr. R. J. J. H. van Son

The research presented in this thesis was part of the project "Functions of Audiovisual Prosody", which was sponsored by the Netherlands Organization for Scientific Research (276-75-001).

Table of contents

Acknowledgements

I could never have written this thesis without the help of so many people, who I would like to thank now.

First, I am very grateful to my supervisors: Marc Swerts and Emiel Krahmer, for the many sparkling discussions in the beginning of the project, for giving me the opportunity to participate in international conferences, for all their comments that significantly improved my writing style, and for their never-ending optimism that this project would come to a successful end.

I also would like to thank the members of the committee: Carlos Gussenhoven, Fons Maes, and Rob van Son, where a special thanks goes to Björn Granström, Bob Ladd and Elizabeth André for traveling all the way to Tilburg.

Carel van Wijk helped me many times with good statistical advice, and Lennard van de Laar provided technical support, and also took care of the lovely cover design. Jean Vroomen was so kind to let us use the experimental software program Pamar. I also must mention the people from the LAU, who were always willing to help with computer problems. Lauraine Sinay and Ingemarie Sam often helped me with practical matters.

I have good memories of the adventure of traveling to Prague for an experiment, which gave me an impression of how exciting research in 'the old days' must have been, and which I could never have accomplished without the help of Graeme van Voorthuijsen, Karel Fliegel, Marie Nilsenová, Prof. Dr. Miloš Klíma and Prof. Dr. Oldřich Uličný.

The chapters that have been published or accepted profited from the comments of Jeffrey Cohn, as well as some anonymous reviewers. Dolf van Enk helped me in locating remaining 'Dunglish' sentences that should have been proper English.

I have a special thanks to all the speakers who allowed me to film their facial expressions. I also must thank the people willing to participate in the experiments.

I would like to thank my roommates with whom I could share many important work-related and private matters: Anja Arts, and Charlotte van Hooijdonk, who is also my paranimf.

I must mention my colleagues and ex-colleagues who made working within the Communication and Information Section very pleasant: Carel van Wijk, Cathy de Waele, Edwin Commandeur, Eric Daamen, Erica Huls, Erwin Marsi, Femke Wieme, Fons Maes, Geno Spoormans, Hans van Driel, Harold Miesen, Ielka van der Sluis, Ingemarie Sam, Jan Renkema, Joost Schilperoord, Juliette Schaafsma, Lauraine Sinay, Lennard van de Laar, Leo Noordman, Leonoor Oversteegen, Lisette Mol, Marie Nilsenová, Marije van Amelsvoort, Pascal Marcelis, Per van der Wijst, Rein Cozijn, Sarah van Vliet, and many others, I have good memories of the walks in the beautiful Warande wood where the squirrels frisk. I also enjoyed having pleasant scientific and personal conversations with fellow PhD-students at the department (and other colleagues), on the way to and from many different occasions, as well as with several people attending the same conferences as I did; I have good memories of practicing the art of 'audiovisual' conversation under different social and cultural circumstances.

I would like to thank my friends, for sharing so many pleasant activities throughout the years, and for your (in some occasions very long) lasting friendship, where you regularly had to cope with my absence due to the writing of this thesis. A very special thanks goes to Irma who is my paranimf. I also want to thank my family for their love and support in all the choices that I make, and my collateral relatives (and step-grandfather) for their interest.

Finally, I express my gratitude to Gerben, for your love, support, humour, inspiration, and for so much that cannot be expressed in words.

When raw wood is carved, it becomes a tool - Lao Tse

1 Introduction

The term "*prosody*" has received a wide range of interpretations in linguistic literature. In a broad sense, prosody can be viewed as the whole gamut of features that do not determine what people are saying, but rather *how* they are saying it. Originally, such a definition was generally used to refer to auditory prosody, i.e. a set of features such as pitch, voice quality, loudness, rhythm, speech rate, and pauses, that are encoded in the speech signal itself (see e.g. 't Hart, Collier, & Cohen, 1990, p.1; Nooteboom, 1997; Rietveld & van Heuven, 2001, pp.231-292). Those features are called "*suprasegmental*", because they "comprise properties of speech that cannot be understood directly from the linear sequence of segments" (van Heuven, 1994, p.2)[1]. Even when some of these prosodic features are typically perceptual in nature (e.g. pitch), they are often expressed in terms of acoustic measures, such as fundamental frequency (F0), duration and amplitude. Also, while some researchers restrict their definition of prosody to a purely phonetic specification of aspects of the speech signal, others include higher-level phonological properties such as intonational phrases, prosodic phrases and metrical feet (Shattuck-Hufnagel & Turk, 1996).

More recently, various researchers tend to broaden its definition to also include visual prosody, i.e. specific forms of body language that communication partners send to each other during the interaction, such as facial expressions, arm and body gestures and pointing (see e.g. Graf, Cosatto, Ström, & Huang, 2002). Both auditory and visual prosody are omnipresent in natural conversations. It would be extremely unnatural to have utterances produced without variations in pitch, tempo, loudness, etc. Similarly, since conversants can see each other in many forms of spoken communication, it would be odd if they were to stay completely immobile during their interactions. Speech, therefore, is multimodal by its very nature. It is likely that not only the production, but also the *perception* of speech, is multimodal. Information from the same distal source will arrive simultaneously through different sensory systems (Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000). For example, as the speaker produces lip movements to create a speech sound, the addressee will receive this information from the visual and the auditory modality (almost) simultaneously through the ear and the eye (Benoît, Martin, Pelachaud, Schomaker, &

---

[1] Individual speech segments consist of vowels, such as an [e], and consonants, such as a [n].

Suhm, 2000; Ghazanfar, Maier, Hoffman, & Logothetis, 2005). To actually use this multimodal information can have several benefits. Input from one modality can replace another one in deteriorated circumstances, such as in a noisy environment (Sumby & Polack, 1954, in Calvert, Brammer, & Iversen, 1998) or in darkness (Calvert et al., 1998). Considering signals from the visual as well as the auditory modality can improve speech perception by a system, or, for example, lip-reading can help the hard of hearing, and signals from two modalities can complement each other, which helps in ambiguous situations (Benoît et al., 2000). However, while we have learned a lot about the pragmatics of auditory prosody, we still miss a good deal of real knowledge of how auditory cues combine with visual ones. It is unclear what the relative importance is of visual cues compared to auditory cues. This will be the main topic in this thesis. We will use the term "*audiovisual prosody*" to refer to the combination of these visual cues with auditory cues.

It is intuitively clear that prosody plays an important role in daily life spoken interactions. In general, it provides utterances with 'extra' information that is often not explicitly contained in the lexical and syntactic make-up of a sentence. Prosody can be used for a wide range of functions, varying from marking the information structure and turn-taking, to adding expressive power, such as emotions and attitudes, to the propositional content of an utterance (see e.g., among others, van Heuven, 1994; Hirschberg, 2002; Rietveld & van Heuven, 2001, p.239 and further).

The current thesis is concerned with a functional analysis of some of the functions of audiovisual prosody. To put our studies into a broader perspective, the rest of this chapter is devoted to a specification of the general research scope, a presentation of our starting assumptions, a review of related studies, and the introduction of our own approach.

1.1 Research scope

When dialogue participants enter a spoken conversation, they will start to establish a *common ground* (Clark, 1996, p.12; Stalnaker, 1978, in Clark & Schaefer, 1989), i.e. a sense of mutual understanding and cooperation. During the entire conversation, speakers and listeners cooperate to ensure that they understand each other well. This process consists of several components. First of all, the dialogue participants determine whether "each utterance is understood as intended" (Clark & Schaefer, 1989, p.261). Listeners provide this *feedback* by sending back-channel signals to the speaker, while the speaker is actively monitoring these signals (Clark & Wilkes-Gibbs, 1986). This is often achieved via an 'opt-out' method: participants display signals when an utterance is *not* correctly understood (e.g. Stalnaker, 1978, in Clark & Schaefer, 1989). Second, participants also engage in other actions during the grounding process, e.g. speakers display signals when they will be finished speaking in order to influence the course of the *turn-taking* (see Sacks, Schegloff, & Jefferson, 1974, for a description of the turn-taking system). Third, participants can display meta-information such as how certain the speaker is about his or her answer (feeling of knowing), what *emotion* (s)he is feeling, or they can use irony or sarcasm (Clark, 1996, pp.110-112, pp.366-374).

In this thesis we will focus upon the role of audiovisual speech within three components of the grounding process: (1) how dialogue participants provide feedback, e.g. how they signal and detect communication problems, (2) how participants regulate turn-taking, e.g. how they display and detect end-of-utterance marking, and (3) how participants display and perceive emotions. In each of these three sub domains we will focus upon the role of the visual and the auditory modality, the relative importance of each modality, and possible interactions between them.

1.2 Starting points

This section describes which starting points underlie the three studies on audiovisual speech in this thesis. We are interested in (1) cross-modal processing, in (2) aspects of the receiver and the sender, and in (3) natural data.

We can illustrate the processes occurring within a dialogue by the well-known Shannon-Weaver model of communication (Weaver & Shannon, 1949). A schematic representation of this model is shown in Figure **1**.

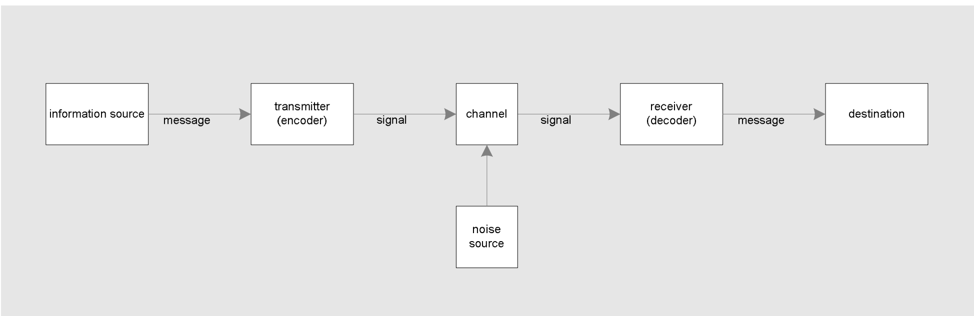

Figure **1** - The Shannon-Weaver communication model (Weaver & Shannon, 1949)

It must be noted that the model was originally designed to describe the signal transmission in telegraphy or in telephone, and it acknowledges the influence of the noise that is possibly exerted upon the signal when sent through the cable (Weaver & Shannon, 1949, pp.33-34).

The model consists of five parts: (1) The information source produces a message out of a set of possible messages, (2) The transmitter (or: encoder) transforms the message into a signal, (3) This signal is sent across the channel to the receiver, (4) The receiver (or: decoder) reconstructs the original message from the signal, and finally (5) the receiver passes the message to the destination, which is the person for whom the message is intended (Weaver & Shannon, 1949, p.7, pp.33-34).

When two persons are involved in a face-to-face dialogue, the cognitive system of the speaker is the *information source*, his or her vocal system is the *transmitter* (or: sender), the air is the *channel* through which the acoustical speech *signal* is transmitted, the ear of the listener is the *receiver*, and his or her cognitive system is the *destination* (Weaver & Shannon, 1949, p.7). Note that when the participants are involved in a telephone conversation or in videoconferencing, or when the conversation is recorded or broadcast, the channel may be different than air, involving technological devices.

1.2.1 Cross-modal processing

When we look at natural, human-human conversations, it becomes apparent that there are important additional aspects that are typical for human communication. To begin with, the signals that are transmitted through a channel do not occur in isolation, but are often accompanied by other signals, either within the same or in another modality.

An auditory speech signal contains prosodic cues, which provide the message with extra meaning. These cues are accompanied by dynamically varying facial expressions, which may also have a prosodic value. In order to visualize the role of the modality, we use a working model which is based upon the original model. Figure **2** shows an adapted version of the Shannon-Weaver model, in which the channel consists of two modalities. Note that the means of the sender are more elaborate than the level of detail shown in the figure: a sender may not only use facial gestures, but also head movements, changes in posture, hand gestures, etc.; a sender may not only use the voice, but also friction in the mouth, clicks, etc. The signal can be sent across the medium as an auditory speech signal, as a visual signal such as a facial expression, or as both.

We want to know what the role is of the visual modality, and whether prosody can be expressed in the visual modality. We are also interested in the relative importance of the visual modality when compared to the auditory modality, as the latter has been investigated more often than the visual modality. Thus, while considering the role of the modality, we want to address several questions: Is 'visual' speech informative, so that it may express a 'visual' equivalent of prosody? If so, is speech coming from different modalities (auditory, visual and the combination of these two) integrated by the receiver, i.e. to what extent do the different modalities complement or obstruct each other? Which modality is the most important?

1.2.2 Receiver versus sender

Another important characteristic of communication that should be captured by the extended communication model, concerns the roles of the sender and receiver. Traditionally, the sender role is associated with the speaker, and the receiver role is associated with the listener. However, recent work (e.g. Clark, 1996) has revealed that this association is too simplistic (see section 1). Not only do the roles switch continuously, but in many cases do dialogue participants play both roles at the same time (a speaker sends information to the listener, but also receives feedback cues indicating whether the information is understood correctly or not). Therefore, we will from now on refer to the sender and the receiver instead of the speaker and the listener, except in those cases where it is obvious who is the speaker and who is the listener. This is represented in the following diagram:

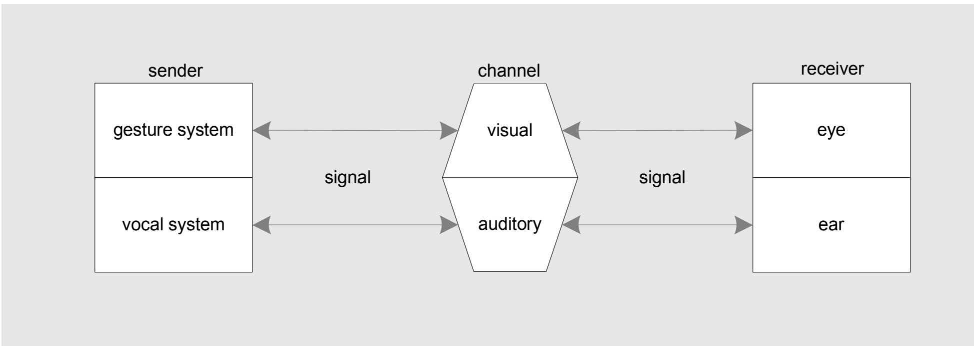

Figure **2** - A communication model of a dialogue

As this thesis addresses the functions of audiovisual prosody in spoken conversations, we will focus on how the signals are sent and received; not on how these signals are subsequently processed in the brain or cognitive system of the receiver and/or sender.

While considering the role of the sender and the receiver, we want to address several questions: Which cues do senders display in audiovisual speech and to which of these available cues are receivers sensitive? Are there individual differences between senders?

1.2.3 Natural data

We investigate sender and receiver behavior in real interactions. Natural audiovisual data is *ecologically valid* because it sheds a light on how audiovisual integration works in daily practice. Our main focus is upon natural spoken conversations, where senders spontaneously display audiovisual signals. In terms of the Shannon-Weaver model, we want the signal to be a natural signal.

Stating that we are interested in natural data is easy. However, obtaining this in practice is more difficult. When *natural expressions* occur in a natural context, the form of these expressions is uncontrolled. It is difficult to instruct dialogue participants to display certain facial expressions (with certain intensity) on command, when they are at the same time involved in a natural conversation. Such an instruction would disturb the conversation and make the displayed expressions less natural. Not only is the form of the expressions uncontrolled, but also the role that they fulfill in the dialogue. It is difficult to determine how the displayed expressions and acoustical signals are linked to the inner state of the person, and therefore what the intention of the person was when displaying the signal. Another problem is that prosodic features can fulfill several functions simultaneously, e.g. a brow raise may indicate "emphasis" or "sadness" (Hirschberg, 2002). It is also uncertain how the signal is perceived by the other dialogue participant. Again, (online) measuring these parameters would disrupt the natural course of the conversation.

Researchers have tried to solve this problem by using actors, who are asked to pose expressions (see e.g. Cohn, Xiao, Moriyama, Ambadar, & Kanade, 2003; Scherer, 2003). But how representative are these *controlled expressions* for natural expressions? Research suggests that controlled visual expressions differ in a fundamental way from natural ones, in that they are more intense and less symmetric in appearance, the latter due to the control by different motor pathways (Zlochower, 2001, and Rinn, 1984, in Cohn et al., 2003), as well as in a number of other ways. An actor may exaggerate the obvious cues and miss the more subtle, natural ones (Scherer, 2003). Another approach is to control the *function* of the expression. An example would be to instruct the participant to answer questions of an interviewer *as if* there was a communication problem. However, this also poses problems, because a participant can express the same message with different prosodic (or linguistic) means, e.g. (s)he can indicate "emphasis" with a brow raise, a louder voice, or by using a different word order (Hirschberg, 2002). More important, a participant is probably unaware *how* (s)he displays these prosodic functions. Further, instructions to display certain behavior on command will also disrupt the conversation. Therefore, one may start to systematically control the *context* in which the expressions occur. The function of the expression can then be (loosely) associated with the role it takes in this particular context, while at the same time the signal is displayed spontaneously. This is a combination of the two previous methods: to *natural expressions* in a *controlled context*.

While considering the naturalness of the data, we want to address the following question: How can we use natural data in an experimental design, so that we can generalize our results to real interactions?

1.3 Types of studies

In this section we describe which methods are currently available in the field of auditory speech as well as in the field of visual expressions.

In the past, different methods have been applied to study expressions while investigating different types of research questions. Traditionally, most research about visual expressions has focused on emotions. Most of these methods have been used to investigate the question of whether "the face reveals emotion in a way that is universally understood" (Fridlund, 1994, p.192). In this thesis, we will focus mainly upon the perceptual value of audiovisual behavior, i.e. the effect that a facial expression (or a voice parameter) can have on dialogue participants, and upon the subsequent course of the dialogue, but not upon cultural or individual differences in expression (for a study of these individual differences see e.g. Cohn, Schmidt, Gross, & Ekman, 2002).

When investigating the perceptual value of facial and vocal behavior, the first step should be the actual *observation* of the behavior. Wagner (1997) has described two approaches in observing behavior (see also Ekman, 1982, pp.46-50):

- *Judgment studies*: The first type of questions considers the communicative function of expressions and therefore how they are interpreted by others (Wagner, 1997). In such studies, we want to know what information is conveyed by the audiovisual behavior and how observers, like dialogue participants, respond to that audiovisual behavior. These studies explore questions like: Can judges detect whether a speaker has reached the end of his or her utterance on the basis of his or her vocal behavior? Or: Can observers (or: judges) tell whether a participant has encountered communication problems in a dialogue by just looking at his or her face?

- *Measurement studies*: The second type of questions addresses the actual facial and vocal behavior itself and thus involves methods that provide a description or measurement of this behavior (Wagner, 1997). These studies explore questions like: Does facial behavior accompanying auditory speech directly reflect the emotional experience of a sender? Or: Are problems during a human-machine dialogue reflected by vocal changes?

Of course, a combination of the two approaches is also possible (Wagner, 1997). According to Wagner (1997), judgment studies are necessary to address questions concerned with the *information* conveyed by an expression. When one wants to reach conclusions about the *components* of the behavior, it is better to use a measurement study.

1.3.1 Judgment studies

Judgment studies have most often been used to study cross-cultural emotion recognition (see e.g. Ekman et al., 1987). Evidence that observers from different cultures are able to see the same emotion(s) in a given face has been presented as support for the assumption that facial behavior has an emotional communicative function (Fridlund, 1994, p.192, see pp.192-268, for a discussion of the cross-cultural studies).

Typically these studies are conducted as follows: The researcher asks members of one culture what face they show in different emotional contexts (e.g. "you feel sad because your child died," "you are angry and about to fight") (Ekman et al., 1987). Thus, the subject is asked to portray emotional *facial expressions*. In the next step, photographs of the portrayals are shown to members of another culture, who have to judge these expressions. Universality is demonstrated when observers in another culture perform better than chance in classifying the photographed facial expressions into the emotional contexts that they are supposed to reflect (Ekman et al., 1987).

Most of these judgment studies were based on static expressions, such as schematic drawings or photographs (see e.g. Ekman, Friesen, & Ellsworth, 1972, pp.49-51; Russell, Bachorowski, & Fernández-Dols, 2003). However, research shows that static and dynamic expressions are processed in a fundamental different way by the brain (Humphreys, Donnelly, & Riddoch, 1993). As we are interested in audiovisual behavior in natural conversations, it is better to use dynamic expressions as stimulus material, because in a natural conversation, the receiver in general performs online processing of the acoustic signal as well as of the constantly moving facial expressions of the sender.

There have been auditory equivalents of the cross-cultural emotion studies, in which *vocal expressions* of emotion are correctly recognized by members of other cultures (Scherer, 2003). The procedure is the same: actors are posing emotions on the basis of a scenario while uttering a standard sentence (Banse & Scherer, 1996). There are differences between emotions according to the modality they are presented in, for example happiness is better recognized in the face while sadness and anger are better recognized in the voice (Scherer, 2003). This is interesting in the light of our assumption considering the role of different modalities.

Other studies focused on non-emotional auditory speech. For example, in a reaction time experiment, subjects had to judge when a turn has ended on the basis of material in which the auditory speech is made unintelligible but is still prosodic (de Ruiter, Miterrer, & Enfield, 2006). In an other study, subjects had to indicate whether they thought that a prosodic boundary would follow after hearing a short sound fragment (Carlson, Hirschberg, & Swerts, 2005). There are, however, few judgment studies in which observers have to judge audiovisual material.

Wagner (1997) describes two types of judgment methods. The first type is the *category judgment method*. Within this approach, the most common method is the *forced choice method*. The stimuli are presented to the judge one at a time, and the judge has to classify the stimulus into a single response category. The other type of method is the *rating method*. The judge has to rate the visibility of a property in the stimulus.

The main advantage of judgment studies is that the stimulus material is investigated free from the context. The judge has no other information than the information which is present in the signal, and can therefore base his or her judgment only upon the features in the signal. If the stimulus is judged correctly, it is possible to establish what characteristics in the stimulus have caused this judgment. Also, it can be established whether the stimulus has a communicative function by itself, and whether this function varies across different judges (for example, judges in different cultures).

On the other hand, the use of judgment studies can pose some methodological problems. It is important to consider these problems during the construction of the experimental design (see Wagner, 1997, p.40, p.46). For example, response biases are common in judgment studies and need to be taken into account. A response bias is the overall tendency of a judge to prefer one answer category above the other. However, response biases can be data themselves, reflecting a factor of interest, and therefore do not a priori pose a problem (Wagner, 1993; Wagner, 1997, pp.46-47).

1.3.2 Measurement studies

Within the methods that try to measure *facial expressions*, different approaches have been taken (Ekman, 1982). Some of these methods try to capture the functions of facial behavior, based on theoretical considerations, and interpret facial expressions in terms of the function of the expression, such as linguistic signals, e.g. the coding of a brow raise as "emphasis", or signals of emotion, e.g. the coding of a brow raise as "sadness" (Bakeman & Gottman, 1997, p.24).

Other methods are based on the anatomical basis of facial action, and can be considered as the more 'objective' coding systems[2]. The facial musculature, which is supposed to be the same for all human beings[3], restricts the number and kind of movements that a face can make (Ekman, 1982; Wagner, 1997). Ekman and Friesen (1978, in Wagner, 1997) applied a technique, used earlier by Duchenne (1862, in Ekman, 1982), of inserting needles which electronically stimulate the (separate) muscles (Ekman, 1979, 1982). Each muscular action received a number, provided that the action caused a visually distinguishable facial movement (Ekman, 1979, 1982). This means that different actions resulting in the same

[2] Of course, there is the factor of the subjectivity of the coder. Judging which muscle caused a visible movement is still a judgment, however, it's a judgment about a perceived *physical* property, a process in which the coders are "supposed to function like machines" (Ekman, 1982, p.48). Calculating the inter-observer reliability will provide a measure of the coders' 'objectivity' and will be an indication of how well the coders succeeded in this attempt.

[3] There are, however, many individual differences in the structure and type of facial muscles, such as the risorius and the zygomaticus major muscle, both believed to play a role in smiling (Schmidt & Cohn, 2001). For example, the zygomaticus major muscle manifests itself as a bifid type in one-third of indiviudals, which is believed to cause cheek dimples (Pessa, 1998, in Schmidt & Cohn, 2001). Also, a large number of individuals lack the risorius muscle (Pessa, 1998, in Schmidt & Cohn, 2001). Further, some people show a specific wrinkle pattern as a sign of sadness instead of raised inner eyebrow corners (Ekman, 2004, pp.99-100).

facial movement are represented by the same number. There are 44 of these distinguishable movements, denoted as Action Units (AU's), which can be produced jointly into more than 7000 different combinations (Ekman, 1982)[4]. The FACS system (Facial Action Scoring System) was designed with the purpose of approximating a comprehensive coding system, i.e. a system which is able to measure the type of all possible actions as well as the intensity and ideally the timing of actions (Ekman, 1982).

The advantage of such a comprehensive coding system is that it is unbiased by an a priori theory (Ekman, 1982). Descriptions of facial behavior are separated from judgments about that same facial behavior, which can then be tested separately, e.g. by the use of judgment studies. Further, comparison of research findings is facilitated by a standardized description of facial behavior (Ekman, 1982). When different researchers use the same method, their results will be easier to compare. The advantage for the researcher when using an existing coding system is that the inter-observer reliability is already established by the training and refinement phase of the system using many different coders.

Disadvantages are that the use of this technique is very time-consuming, especially when applied to dynamic facial behavior, demanding a slow-motion replay of film fragments (Bakeman & Gottman, 1997, p.25). Questions can then be raised about the external validity of facial expressions when scored by such a slow-motion replay. Another disadvantage is that a full coding of all facial actions yields a considerable overhead, as it is likely that not all facial actions are equally relevant for the particular research question. It can save time to measure only the features hypothesized to be actually used by the receiver, instead of measuring all features.

Therefore, it may be better to select only a few features which are suspected to be the most useful for a particular situation and to develop a coding system that is more in tune with the research question (Bakeman & Gottman, 1997, p.25). We will refer to this method as *restricted annotation*. One way to make this reliable is by using the inter-observer agreement (e.g. using the kappa-statistic) when coding individual features (Carletta, 1996; Scherer, 2003; Wagner, 1997). This method is realized as follows: A number of observers look for specific features in the data, and whenever a feature is encountered it is marked on a binary or gradual scale. Their individual scores are compared statistically. If the correlation

---

[4] Later, 3 of these Action Units are merged into one. These are the AU's 25 to 27, as they concern the opening of the mouth by dropping the lower jaw, which can be regarded as the same movement but with a different intensity. Also, AU 41, 42 and 43 are now coded according to criteria of intensity; see also http://www-2.cs.cmu.edu/afs/cs/project/face/www/facs.htm.

is high enough, the feature is tagged as "present" in the stimulus. These features can then be compared with the judgment scores in a perception task (Scherer, 2003). The correlations will tell whether the present features display the functions under investigation.

An alternative for a feature-based coding system is a system based upon the xy-dimensions in the visual signal (McNeill, 1992, pp.377-387; McNeill, 2005, pp.273-275). McNeill divides the gestural space into two dimensions (horizontal and vertical), in which a number of squares are located. According to the viewpoint of the speaker, each gesture is located at one of the squares in this space, e.g. in the center, or in the upper left of the periphery. Apart from the location, the form and meaning of the gestures are also coded.

In the *auditory domain*, several measurement techniques are available. Again, a division is made between describing the form of the acoustical signal and the function. The prosodic form can consist of rising and falling pitch, longer and shorter segments, and more (Hirst, 2004). Software which can extract prosodic information from the auditory speech signal is generally available (e.g. PRAAT), using algorithms producing a phonetic figure of the fundamental frequency curve (Boersma & Weenink, 2007; Hirst, 2004). Such prosodic patterns can be further described using more abstract annotation systems, such as the ToBi system (Hirschberg, 2002; Hirst, 2004). The ToBi system (Tones And Break Indices) represents the pitch accents that have been described according to a prosodic model of American English, with set of discrete symbols (Hirschberg, 2002; Hirst, 2004).

The advantage of using the ToBi system is, among other advantages, that it allows researchers to share their findings (Hirschberg, 2002).

Again, the main disadvantage is that transcribing can be time-consuming, and alike the visual annotation systems, it can be wise to use only a part of the annotation system (Hirst, 2004). Also, the inter-observer agreement is not high for determining the type of pitch accent or boundary (Hirst, 2004).

1.4 Approach in this thesis

In this section we describe how we developed a method on the basis of our starting points, suitable for studies on the *combination* of auditory speech and facial expressions. The central characteristic of the paradigm that we have developed is the combination of the elicitation of audiovisual speech with experiments that tests the perception of that audiovisual speech. We will illustrate how the paradigm can be put into practice in the three studies described in Chapters 2 to 5.

The models that we have discussed can be divided into measurement and judgment methods, and explore sender as well as receiver behavior. A very interesting possibility is to combine these methods in order to investigate the relationship between actual audiovisual behavior and the message that this same behavior transmits. We use an approach, in which we first elicit audiovisual speech, and next we test the perception of that audiovisual speech (see also Carlson & Swerts, 2003; Krahmer, Swerts, Theune, & Wegels, 2002). According to Scherer (2003) there are three types of studies: encoding studies, in which expressions are elicited, decoding studies (=judgment studies), in which the perception of the expressions is tested, and inference studies (=measurement studies), in which the underlying cues responsible for the receiver's inferences are investigated. Following his method, our studies consist of three phases:

- We *elicit audiovisual expressions* in a semi-spontaneous, but highly controlled way, so that we know exactly what the context of these expressions was (elicitation).
- Next, we establish the conditions under which the perception of specific audiovisual behavior occurs using a *judgment* method (perception test).
- And finally, we describe the audiovisual behavior which accompanied these conditions using a *measurement* method (measurement of the signal recorded during the elicitation).

In the first place we are interested in how human receivers actually perceive audiovisual speech, but later we will investigate which cues caused that perception.

Our approach meets the three starting points that were described in the introduction (see section 1.2):

- *First, the elicited audiovisual speech is recorded and can be presented in several modalities.* By systematically varying parameters in the presentation, it can be established what the relative importance is of the different levels of the parameter in the perception. For example, because we are interested in the integration and interaction of the different modalities, we do not just present the stimuli the way they were recorded, but we manipulate the recorded stimulus material with respect to the different *modalities*, by creating a unimodal condition: vision-only (VO), or audio-only (AO), or by using the original, bimodal material: audiovisual (AV).

- *Second, the elicited speech can be used as stimulus material for perception tests.* Because we elicit the utterances in a controlled way, where the content as well as the order of the dialogue is controlled, it becomes possible to investigate how these utterances are actually perceived by receivers. Thus, we investigate the relationship between the systematically varied context in which the original behavior is displayed - e.g. during communication problems, a complete utterance within a sender's turn, or in the case of an elicited emotion - and what part of that experimental context can be recovered using only a part of the displayed facial (or vocal) behavior as the only available information source - e.g. when listening to a system's question, a single word, or an emotional sentence.
- *Third, the approach enables us to elicit natural (albeit controlled) audiovisual speech.* The conditions in which the audiovisual speech is elicited are highly controlled, producing natural (but controlled) audiovisual speech. This means that we first record participants engaged in a natural interaction. In this natural interaction, active participation may take place, e.g. the participant may answer a question, but the participant can also listen to the other (non-human or human) dialogue participant, e.g. when a dialogue system asks a question. All these receiver and sender behaviors of the participant are recorded. Therefore, the stimulus material resembles the signals displayed in natural interactions.

1.4.1 Overview

The remainder of this thesis is structured as follows.

Chapter 2 describes studies carried out to investigate how communication problems are reflected in audiovisual speech during a human-machine dialogue. These studies make use of audiovisual recordings of an interaction of a user with a spoken dialogue system, which contain samples of problematic moments in the dialogue. In three perception tests, using different types of samples presented in different modalities, participants have to classify these recordings as problematic or non-problematic. In an additional observational analysis, the results of these perception tests are linked to visual features in the stimuli, in order to find out which features are potential cues for error detection.

Chapter 3 investigates how speakers approaching the end of their utterance reflect this in their audiovisual speech, which may play a role in the fluency of turn-taking. These studies make use of audiovisual recordings of participants producing lists of words in an interview setting, providing samples of non-final and final moments. In a reaction time experiment, using different modalities, participants have to indicate when the end of an utterance in a recording is reached. In a second perception test, using different modalities, participants have to classify the samples as final or non-final. In an additional observational analysis, the results of these perception tests are linked to visual features in the stimuli, in order to find out which features are potential cues for end-of-utterance detection.

Chapter 4 investigates how speakers display audiovisual emotional speech. This study makes use of audiovisual recordings of participants displaying positive and negative emotions invoked via a Dutch variant of the Velten method. These emotions can be congruent or incongruent with the (emotional) lexical content of the uttered sentence. In a perception test, using different modalities, Czech participants have to rate the perceived emotional content of the recordings. In a second perception test, using a gating paradigm, Dutch participants have to classify the recordings, which are presented in only the visual modality, as positive or negative. In an additional observational analysis, the results of these perception tests are linked to visual features in the stimuli, in order to find out which features are potential cues for emotion perception.

Chapter 5 presents an overview of the main results and general conclusions will be drawn. Also, limitations of this study are discussed and directions for future research will be suggested.

2 Communication problems

in human-machine interactions

2.1 Introduction

The goal of the investigation presented in this chapter is to explore to what extent it could be beneficial to use features of a user's facial expression to detect communication problems in his or her interactions with a spoken dialogue system[5].

It is well-known that managing communication problems in spoken human-computer interaction is difficult. One key issue is that spoken dialogue systems are not good at determining whether the communication is going well or whether communication problems arose (e.g. due to poor speech recognition or false default assumptions). The occurrence of problems negatively affects user satisfaction (Walker, Litman, Kamm, & Abella, 1998), but also has an impact on the way users communicate with the system in subsequent turns, both in terms of their language and speech. For instance, when users notice that a system has difficulties to handle their prior spoken input, they tend to produce utterances with marked linguistic features (e.g. longer sentences, marked word order, more repeated information, etc.) (Krahmer et al., 2002). In addition, human speakers respond in a different vocal style to problematic system prompts than to unproblematic ones: when speech recognition errors occur, they tend to correct these in a *hyperarticulated* manner (which may be characterized as longer, louder and higher). This generally leads to worse recognition results (*spiral errors*), since the standard speech recognizers are trained on normal, non-hyperarticulated speech (Hirschberg, Litman, & Swerts, 2004; Levow, 2002; Oviatt, MacEachern, & Levow, 1998), although more recent studies suggest that systems become less vulnerable to hyperarticulation (Goldberg, Ostendorf, & Kirschhoff, 2003). In a similar vein, when speakers respond to a problematic yes-no question, their denials ("no") share many of the properties typical of hyperarticulated speech, in that they are longer, louder and higher than unproblematic negations (Krahmer et al., 2002).

---

[5] An earlier version of this chapter was published in Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication*, *45*(3), 343-359.

In other words, one could state that dialogue problems lead to a marked interaction style of users, which manifests itself partly in a set of prosodic correlates. Based on these observations, it has been suggested that monitoring prosodic aspects of a speaker's utterances may be useful for problem detection in spoken dialogue systems. It has indeed been found that using automatically extracted prosodic features helps for problem detection (Hirschberg et al., 2004; Lendvai, van den Bosch, Krahmer, & Swerts, 2002). While this has led to some improvements, the extent to which prosody is beneficial differs across studies. Moreover, in all these studies a sizeable number of problems are not detected. In general, it appears that the detection of errors improves if prosodic features are used in combination with other features already available to the system, such as more traditional acoustic or semantic confidence scores, knowledge about the dialogue history, or the grammar being used in a particular dialogue state (Ahrenberg, Jönsson, & Thurée, 1993; Bouwman, Sturm, & Boves, 1999; Danieli, 1996; Hirschberg, Litman, & Swerts, 2001; Litman, Hirschberg, & Swerts, 2000). The current chapter explores whether it is potentially useful to include yet another set of features, i.e. visual features from the face of the user who is interacting with the computer.

Indeed, it makes sense to assume that a speaker's facial expressions may signal communication problems as well. One obvious reason is that hyperarticulation is likely to be detectable from inspecting more exaggerated movements of the articulators. Erickson, Fujimura & Pardo (1998) found that speakers' repeated attempts to correct another person are highly correlated with more pronounced jaw movements, which are likely to be clearly visible to their addressees (see also Dohen, Lœvenbruck, Cathiard, & Schwartz, 2004; or Gagné, Rochette, & Charest, 2002 about related visual correlates of contrastive stress). In addition, in line with the earlier observation that speakers adapt their language and speech after communication errors to a more marked interaction style, there is evidence that speakers also change their facial expressions in problematic dialogue situations. Swerts, Krahmer, Barkhuysen & van de Laar (2003) applied the so-called *feeling-of-knowing* paradigm (Brennan & Williams, 1995; Hart, 1965; Smith & Clark, 1993) to investigate how speakers cue that they are certain or rather uncertain about a response they give to a general factual question. It was found that it is indeed often clearly visible when people were insecure about the answer to a response, in that speakers show much more deviations from "normal" facial expressions (e.g. more eyebrow movements and gaze acts). Given such observations, it is worthwhile to investigate whether speakers also exhibit special visual expressions when they are confronted with communication problems in spoken human-machine interactions.

This research fits in a recent interest to try and integrate functional aspects of facial expressions in multi-modal systems, with the ultimate goal to make the interaction with such systems more natural and efficient. Some systems already supplement their interface with an Embodied Conversational Agent (ECA), for instance in the form of a synthetic head, to support the communication process with users. Visual cues of such ECA's appear to be functionally relevant in more than one respect. They make the speech more intelligible (e.g. Agelfors et al., 1998; see also Jordan & Sergeant, 2000), and can give clues about the status of the information a system sends to the user, for instance to signal the difference between negative or positive feedback responses from a system (Granström, House, & Swerts, 2002). An additional advantage of using a synthetic face is that it can give silent cues about the internal state of the system, e.g. to signal that it is paying attention to the user or that it is looking for information, following the general best practice to make a system's behavior and reasoning clear to a user (Sengers, 1999).

The perspective in the current chapter is different from that of such earlier studies in that it does not concentrate on multi-modal features of system utterances, but rather deals with analyses of the users' facial expressions. The exploitation of the users' auditory *and* visual cues is becoming a real possibility in advanced multi-modal spoken dialogue systems (see e.g. Benoît et al., 2000), which combine speech recognition with facial tracking.

Earlier work in bimodal speech recognition has shown that using automatic lip-reading in combination with more standard automatic speech recognition techniques leads to a reduction of the number of recognition errors (see e.g. Petajan, 1985). In addition, comparable to the silent visual cues from a system, facial expressions of a user may indicate communication problems even when the person is *not* speaking, but for instance when (s)he becomes aware of a communication problem during the system's feedback. Such cues clearly have added value compared to the auditory and linguistic cues to errors used before, because they would enable a very early detection of problems. Obviously, this would be useful from a system's point of view, since the sooner a problem can be detected, the earlier a repair strategy may be started (e.g. a re-ranking of recognition hypotheses or a modification of the dialogue strategy).

Therefore, the general goal of the research described in this chapter is to investigate the information value of a speaker's visual cues for problem detection in spoken human-machine interaction.

The study consists of two parts. First, we describe three perception experiments in which participants were shown selected recordings of Dutch speakers engaged in a telephone conversation with a train timetable information system[6]. The recordings constituted minimal pairs as they were very comparable in terms of their words and syntactic structure but differed in that they were excised from a context which was either problematic or not. The recordings were presented without the original context to participants who had to determine whether the preceding speaker's utterance had led to a communication problem or not. The first experiment focuses on participants' responses during verification questions of the system (i.e. when participants listen in silence), which either verify correct or misrecognized information. The second experiment concentrates on speakers uttering "no", either in response to a problematic or an unproblematic yes-no question from the system. The third experiment, finally, is devoted to speakers uttering a destination station (filling a slot), either for the first time (no problem) or as a correction (following a recognition error). The descriptions of these three studies are preceded by an overview of the general experimental procedure.

Second, we describe the results of some observational analyses. We attempt to find visual correlates of problematic situations that could have functioned as cues to participants in the different perception studies described in section 2.3.

Our major finding is that more problematic contexts lead to more dynamic facial expressions, in line with earlier claims that communication errors lead to marked speaker behavior. We conclude this chapter with a general discussion and some perspectives on further research.

2.2 Audiovisual recordings

The stimuli used in the three experiments were all taken from an audiovisual corpus of speakers engaged in telephone conversations with a speaker independent Dutch spoken dialogue system providing train timetable information. The original corpus consists of 9 speakers (5 male and 4 female) who query the system on 7 train journeys (63 dialogues in total). Each dialogue took approximately 25 minutes. In 76% of the dialogues speakers finish the task successfully (i.e. they obtain the correct advice).

---

[6] In the remainder, the term "speakers" refers to users who were recorded while they interact with a spoken dialogue system.

The original recordings were made with a digital video camera (25 frames per second). Speakers were led to believe they were involved in the data collection required for a new kind of "video-phone", hence they were instructed to face the camera at all times. Also, to ensure an optimal view of the face without a phone device blocking important visual features, speakers had to interact via a mobile phone positioned in front of them on a table. Afterwards the recordings were read into a computer and transcribed. On the basis of the transcriptions it could be decided which speaker utterances were misrecognized or misunderstood, and thus led to communication problems. It turned out that 374 out of 1183 speaker turns were misunderstood by the system (32%). These figures are representative of speaker independent spoken dialogue systems in real life settings (e.g. Carpenter et al., 2001; Hirschberg et al., 2004; Nakano & Hazen, 2003; Walker et al., 1998).

2.3 Perception studies

2.3.1 Stimuli

For all three perception studies, the stimuli (verification questions, negations and slot-fillers respectively) were randomly selected on the basis of the transcribed dialogues. Per speaker, two problematic and two unproblematic instances were selected. If this turned out to be impossible for a speaker, that speaker was omitted from the experiment. Therefore, in the second study, only 7 speakers were selected from the corpus, and in the third study, 8 speakers were selected. In the perception studies, the stimuli were always presented per speaker and in a random order. Each block of four stimuli per speaker (two problems, two non-problems) was preceded by a reference stimulus showing that speaker in an unproblematic situation. Each study started with a short exercise session containing two problematic and two unproblematic stimuli (and a reference stimulus), in order to make participants familiar with the kind of stimuli and the experimental setting. See Figure **3** for two representative illustrations of speaker ED.

2.3.2 Design

The experiment had a within-subjects design, with the factors speaker (speaker 1 to 8) x problem (with levels NON-PROBLEMATIC and PROBLEMATIC) x two instances. All stimuli were presented in one, AUDIOVISUAL (AV), condition. The dependent variable is the percentage of participants that classify the fragment as problematic.

Figure **3** - Two stills from speaker ED uttering the phrase "nee" (no) in an unproblematic (left) and a problematic situation (right)

2.3.3 Participants

A group of 66 participants (20 male and 46 female, all students from Tilburg University) participated in the three experiments, all but one native speakers of Dutch. The participants were between 19 and 47 years old.

2.4 Experiment 1: System questions

2.4.1 Procedure

In the first study, participants saw speakers listening to verification questions. These verification questions can be unproblematic, such as the system question in example (1).

(1)	User:	Amsterdam.
	System:	So you want to travel to Amsterdam?

But they can also verify misrecognized information as in (2):

(2)	User:	Rotterdam.
	System:	So you want to travel to Amsterdam?

In the first study, participants have to determine on the basis of the speaker's facial expressions during the system's explicit verification questions, whether the verified information is correct (as in 1) or not (as in 2). They were shown 4 verification questions for all 9 speakers (36 stimuli in sum). For each speaker, two verification questions followed a recognition error and two did not.

2.4.2 Statistical analyses

All tests for significance were performed using a X^2 test.

2.4.3 Results

The results are presented in Table **1**. Inspection of the table reveals that most speakers' reactions to unproblematic verification questions are indeed classified as unproblematic by the majority of the participants. The overall mean of participants who perceive unproblematic stimuli as *problematic* is only 26%. On the other hand, most participants indeed classify speakers' reactions to problematic verification questions as signals of a problem (overall mean 75%).

Table **1** - Percentage of participants who classify an instance of a speaker listening to a system's utterance as a signal of a problem

Speaker	*¬P1*	*¬P2*	*P1*	*P2*
AA	$.00^{c}$	$.01^{c}$	$.73^{c}$	$.94^{c}$
CH	$.80^{c}$	$.20^{c}$	$.99^{c}$	$.99^{c}$
DB	$.24^{c}$	$.30^{b}$	$.94^{c}$	.50
EC	$.20^{c}$	$.00^{c}$	$.62^{a}$	.59
ED	.61	.58	$.97^{c}$	1.0^{c}
IB	$.03^{c}$	$.23^{c}$	$.36^{a}$	.56
LS	$.28^{c}$	.53	$.94^{c}$	$.29^{c}$
PM	$.20^{c}$	.46	$.99^{c}$	$.38^{a}$
SB	$.06^{c}$	$.03^{c}$	$.88^{c}$	$.99^{c}$
Mean	.26		.75	

† ^{a} = $p < .05$; ^{b} = $p < .01$; ^{c} = $p < .001$. For 9 speakers, participants classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2)

Table **2** summarizes the classifications from Table **1**: for 12 of the 18 problematic verification questions and for 13 of the 18 unproblematic ones did a statistically significant number of participants make the correct classification. Note that some of the stimuli were systematically misclassified (in particular, utterance ¬P1 of speaker CH, utterance P1 of speaker IB, utterance P2 of speaker LS and utterance P2 of speaker PM).

Table **2** - Contingency table summarizing the number of significant classifications (problem and ¬problem) for the different conditions (problem and ¬problem) from Table **1**, non-significant classifications are counted as random

Condition	*Problem*	*¬Problem*	*Random*	*Total*
Problem	12	3	3	18
¬Problem	1	13	4	18
Total	13	16	7	36

2.4.4 Summary

The results of the first study show that participants are generally capable of correctly determining whether a verification question contained a problem or not, solely on the basis of a speaker's facial expression during the verification. This shows that keeping track of facial expressions during spoken human–machine interactions can be helpful, even when speakers are silent. Closer inspection of the stimuli suggests that during unproblematic verification questions, participants maintain a neutral facial expression throughout, while they become more expressive (e.g. moving, laughing or frowning) during problematic verification questions. Interestingly, the aforementioned systematic misclassifications support this informal observation, in that speaker CH frowns during an unproblematic system question, while speakers IB, LS and PM keep a neutral expression during a system question which verifies misrecognized information. PM differed from the other two speakers in the sense that he also smiled in the film fragment.

2.5 Experiment 2: Negations

2.5.1 Procedure

In the second study, participants saw speakers only uttering a negation ("nee", *no*). This could be a response to a yes-no question which does not verify recognized information (so speakers by definition do not become aware of a communication problem), but instead offers the speaker a choice in the possible course of action taken by the system in the subsequent dialogue, as in example (3):

(3)	System:	Do you want me to repeat the connection?
	User:	No.

On the other hand, if the question verifies a misrecognition (cf. example (2) above), participants' "no" signals a communication problem:

(4)	System:	So you want to travel to Amsterdam?
	User:	No.

Participants of the perception study saw only the "no" utterances, presented without any further context, and had to determine whether the speaker signaled a communication problem (as in 4) or not (as in 3). Stimuli from seven speakers were used in the second study, with a total of 28 negations. Two speakers were omitted, as it was not possible to obtain a balanced set from their data.

2.5.2 Statistical analyses

All tests for significance were performed using a X^2 test.

2.5.3 Results

The results of the second study can be found in Table **3**. The results show that participants found this test much harder than the first one.

Overall, the unproblematic negations are perceived as problem indicators by 41% of the participants, while the problematic ones are perceived as signaling a problem by 52% as the participants. Clear differences between speakers exist. Speaker LS is often misclassified: the two unproblematic utterances are both significantly classified as signals of a problem, while the two problematic utterances score random (most participants consider them unproblematic).

Table **3** - Percentage of participants who classify a "no" utterance as a signal of a problem

Speaker	*¬P1*	*¬P2*	*P1*	*P2*
AA	.49	.27[c]	.59	.50
CH	.08[c]	.26[c]	.76[c]	.53
EC	.59	.58	.41	.39
ED	.39	.46	.88[c]	.68[b]
IB	.18[c]	.52	.18[c]	.65[a]
LS	.71[c]	.68[b]	.45	.42
SB	.38[a]	.27[c]	.24[c]	.70[c]
Mean	.41		.52	

† [a] = $p < .05$; [b] = $p < .01$; [c] = $p < .001$. For 7 speakers, participants classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2)

Closer inspection of the stimuli reveals that LS was frowning in the unproblematic utterances. Overall, in about half of the cases no significant preference in either direction exists (see Table **4**). Of the 15 stimuli for which the classification showed a significant pattern, the majority is in the expected direction.

The significant misclassifications for the unproblematic cases are both due to LS. The significant misclassifications for the problematic cases are due to IB and SB. A first inspection of their recordings shows that IB displayed little or no facial expressions, while SB showed strong head movements and was nodding.

Table **4** - Contingency table summarizing the number of significant classifications (problem and ¬problem) for the different conditions (problem and ¬problem) from Table **3**, non-significant classifications are counted as random

Condition	*Problem*	*¬Problem*	*Random*	*Total*
Problem	5	2	7	14
¬Problem	2	6	6	14
Total	7	8	13	28

2.5.4 Summary

In general participants found it difficult to determine on the basis of just the "no" whether this negation signaled a communication problem or not. In roughly half of the cases, there was no significant tendency in either direction. Of the remaining cases most of the classifications were correct. This outcome weakly confirms earlier work on the perception of negations (Krahmer et al., 2002); albeit that participants had more difficulty in classifying the negations in the current experiment.

This could be due to the fact that the negation phrases in Krahmer, Swerts et al. (2002) were always cut from longer utterances (e.g. "no, thanks" or "no, to Rotterdam!"). Alternatively, it could also be that the visual modality distracts listeners from the prosodic cues (compare Doherty-Sneddon, Bonner, & Bruce, 2001). Also the unproblematic negations occurred always at the end of the original conversation, so it may have been possible that the speakers' faces showed irritation after being misunderstood earlier in the conversation.

2.6 Experiment 3: Destinations

2.6.1 Procedure

In the third study, participants saw speakers uttering a destination. This could be in a no-problem context like (5):

(5)	System:	To which station do you want to travel?
	User:	Rotterdam.

Or, it could be a correction in response to a verification question of misrecognized or misunderstood information (cf. (2) above):

(6)	System:	So you want to travel to Amsterdam?
	User:	Rotterdam.

For the third study 8 speakers were selected, with a total of 32 stimuli. One speaker was omitted, as it was not possible to obtain two problematic and two unproblematic stimuli from his dialogues.

2.6.2 Statistical analyses

All tests for significance were performed using a X^2 test.

2.6.3 Results

Table **5** displays the results per speaker, and Table **6** summarizes these results. The overall results are closely related to those of the first study: most participants classify most non-problematic destinations as unproblematic, and they classify most problematic destinations as problematic. Again differences between speakers are found, most notable here is that 4 unproblematic slot-fillers are significantly classified as problematic. An inspection of these film fragments show that some of the speakers were frowning, and all were hyperarticulating.

Table **5** - Percentage of participants who classify an instance of a speaker uttering a destination as a signal of a problem

Speaker	*¬P1*	*¬P2*	*P1*	*P2*
AA	$.68^{b}$	.53	$.73^{c}$	$.65^{a}$
CH	$.14^{c}$	$.67^{b}$	.61	$.94^{c}$
DB	$.11^{c}$	.47	$.99^{c}$	$.97^{c}$
EC	.53	$.70^{b}$	$.00^{c}$	.39
ED	.61	$.70^{b}$	.61	1.0^{c}
IB	$.05^{c}$	$.26^{c}$	$.99^{c}$	$.80^{c}$
LS	$.06^{c}$	$.26^{c}$	.56	$.70^{b}$
SB	$.20^{c}$	$.32^{b}$	$.79^{c}$	1.0^{c}
Mean	.39		.73	

† ^{a} = $p < .05$; ^{b} = $p < .01$; ^{c} = $p < .001$. For 8 speakers, participants classified two non-problematic stimuli (¬P1 and ¬P2) and two problematic ones (P1 and P2)

Another striking outlier is utterance P1 from EC, which all 66 participants classified as unproblematic. The fragment shows that this speaker displayed a single head movement, but no further movements.

Table **6** - Contingency table summarizing the number of significant classifications (problem and ¬problem) for the different conditions (problem and ¬problem) from Table **5**, non-significant classifications are counted as random

Condition	*Problem*	*¬Problem*	*Random*	*Total*
Problem	11	1	4	16
¬Problem	4	8	4	16
Total	15	9	8	32

2.6.4 Summary

In a majority of cases participants were capable to correctly classify speaker's utterances of destinations. Inspection of the stimuli suggests the same basic picture as for the first study: when there are no problems, participants have a neutral facial expression, when they need to correct misrecognized information they become more expressive. Audiovisual hyperarticulation appears to be a clear cue for this.

2.7 Observational analysis

2.7.1 Introduction

The series of perception experiments described above brought to light that participants are generally capable to detect problematic dialogue events on the basis of observations of recorded film fragments of human-machine interactions. While participants also had access to possible speech cues in the video films, there are reasons to believe that visual signals have undoubtedly played a role as well in their classification of problematic and unproblematic events. In particular, since the speakers did not talk at all in experiment 1, participants could only have paid attention to facial expressions from the recorded speaker.

To gain further insight into such visual cues, we annotated all fragments in terms of a number of facial features that could have functioned as cues to problematic or unproblematic dialogue events.

In the next sections, we will first describe the labeling procedure we defined, and then embark on the results of analyses where we correlate the annotated features both with the actual and the perceived problems described in the earlier part of the chapter. It will be shown that problematic dialogue sequences are characterized by more dynamically varying facial expressions of users, in line with earlier observations that speakers switch to a marked interaction style in terms of their language and speech in the case of problems.

2.7.2 Labeling

In order to determine which visual cues influenced participants' judgments we labeled the fragments mentioned above using a set of facial features. The choice of these features was primarily based on the results of pilot observations of a subset of the recorded video fragments (see various discussion sections above). The labels consist of seven different visual features, five of which are defined and visualized in Table **7**. The chosen features are roughly comparable with Action Units (AU's) described by Ekman and Friesen (1978), though there is not necessarily a one-to-one mapping to these Action Units. These Action Units constitute the basic ingredients for the influential Facial Action Coding System (FACS) which assumes that every visible facial movement is the result of muscular action. Therefore, a comprehensive coding system can be obtained by discovering how each muscle of the face acts to change a unique visible appearance. With that knowledge it would be possible to analyze any facial movement into anatomically based uniquely discriminable Action Units. Table **7** in particular displays examples of marked settings of SMILING (AU 12-13), DIVERTED HEAD POSITION (AU 51-58), EYE MOVEMENTS (AU 61-64), FROWNING (AU 4) and EYEBROW RAISING (AU 1-2). Additional visual features not shown in this table are FINAL MOUTH OPENING (AU 25-27) (i.e. whether a speaker silently opened his mouth at the end of the video film to prepare for upcoming speech) and the occurrence of (vertical or horizontal) REPETITIVE HEAD GESTURES (basically reflecting a "yes" or a "no" signal); both are difficult to visualize using a single still image. All of these features were labeled as discrete events, in terms of presence or absence of a marked setting of the feature, except for DIVERTED HEAD POSITION and SMILING which were given a number on a small scale between 0 and 2 to reflect different strengths, where 0 stands for a complete absence and 2 represents a very clear presence of a diverted head position or smiling.

The repetitive head gestures, when present, were given a different label according to whether they represented a vertical ("yes") or horizontal ("no") gesture. In addition to these purely visual features, we also included one primarily auditory one, i.e. the occurrence of HYPERARTICULATION. The presence of hyperarticulation was largely determined on the labelers' auditory impression of whether the speech was generally spoken with a louder voice, higher pitch, and/or at a slower rate, though it is clear, as already suggested by earlier findings of Erickson et al. (1998), that hyperarticulation was also cued visually. Following procedures outlined by Wade, Shriberg & Price (1992), HYPERARTICULATION was given a number between 0 and 2 to distinguish different degrees of hyperarticulation, where 0 represents complete absence and 2 a very strong form of hyperarticulation.

Table **7** - Selection of a number of annotated features; the example and the description represent the marked settings for each feature

Label	*Example*	*Description*
Smiling		Speaker produces a clearly visible smile or laughter
Diverted head position		Speaker moves head away from its position at onset
Eye movements		Speaker diverts eye gaze from its position at onset, relative to the position of the head
Frowning		Speaker produces a frown, primarily visible in the forehead or between the eyebrows
Eyebrow raising		Speaker raises one or two eyebrows from neutral position

The labeling was performed by the three authors of this chapter. The procedure was as follows. The coders watched the film fragments and labeled them using a set of eight features, i.e. the seven visual features plus HYPERARTICULATION. Each coder labeled each feature individually. Comparing the labelers' individual scores showed an agreement in most of the cases (80%), where agreement is computed by counting the number of video fragments which received total consensus (three identical annotations for all eight features) divided by the total number of fragments. If a feature was labeled on a scale and the individual scores on the scale did not match (e.g. one coder saw minor hyperarticulation ('1') and the two other coders noted very clear hyperarticulation ('2'), this was also regarded as disagreement. The film fragments of the destinations invoked the largest amount of disagreement (25%). The features upon which there was most disagreement were: HYPERARTICULATION (48%) and DIVERTED HEAD POSITION (38%), whereas coders always agreed on the annotation of FINAL MOUTH OPENING. One complicating factor in the labeling process was that the different features are not entirely independent and are sometimes difficult to separate, such as the potential co-occurrence of a single head movement (DIVERTED HEAD POSITION) and REPETITIVE HEAD GESTURES which could result in nodding. Also, it was not always obvious to determine whether the face varied in terms of a head movement alone, or in combination with diverted gaze. For the analyses below, disagreements between labelers were resolved via majority voting for the discrete features, while the scores for the continuous features (DIVERTED HEAD POSITION, SMILING and HYPERARTICULATION) were summed resulting in an overall score between 0 and 6 for these respective features.

2.7.3 Results

In the results section, we explore to what extent there is a relation between the *perceived* problems in the three experiments and the annotated audiovisual features described above. In addition, we also investigate the relation between the audiovisual cues and the *actual* presence or absence of problems in the stimuli.

Audiovisual features and the perception of problems

First, we will look at various correlations of these features with the proportion of participants who classify a film fragment as problematic. To this end, we will take a purely perception-oriented approach, in the sense that we do not take into account whether or not the fragment was originally extracted from a problematic or unproblematic dialogue context. In other words, what matters is how that fragment is classified by a subject, irrespective of whether that classification was correct or not. The results are shown in Table **8**, which gives the overall results for the stimuli used in experiments 1-3, respectively. HYPERARTICULATION does not play a role in experiment 1 (the speaker silently listens to the system), and is treated as a missing value in that experiment. For the purpose of simplicity we recoded the scalar features to binary ones in this table (but see below). The results are presented in the form of different 2-by-2 matrices, which give the distributions of utterances perceived as problematic or not problematic as a function of the presence or absence of a marked feature setting. The significance and the strength of the associations are expressed in terms of χ^2 and Cramer's *V* tests, respectively.

Table **8** - Distribution of utterances from experiments 1-3 perceived as problematic or not problematic as a function of the presence or absence of a marked feature setting

Feature	*Present*	*Perceived as*		*Statistics*	
		¬Problem	*Problem*	X^2	*Cramer's V*
Hyperarticulation	No	854	466	221.7[a]	.237
	Yes	1046	1594		
Smiling	No	2455	2231	119.0[a]	.137
	Yes	607	1043		
Diverted head position	No	1015	1097	.1	.004
	Yes	2047	2177		
Frowning	No	2539	1883	484.4[a]	.277
	Yes	523	1391		
Eyebrow raising	No	2665	2615	58.4[a]	.096
	Yes	397	659		
Eye movements	No	1671	1563	29.6[a]	.068
	Yes	1391	1711		
Mouth opening	No	2256	2628	38.9[a]	.078
	Yes	806	646		
Repeated head gestures	No	2452	2762	99.4[a]	.125
	Horiz	144	252		
	Vert	466	260		

† [a] = $p < .001$. The significance and the strength of the associations are expressed in terms of $\chi 2$ (*df* = 1, except for repeated head gestures where *df* = 2) and Cramer's *V* tests, respectively

The overall results show that almost all features had a significant impact on the way an utterance is perceived as problematic or not: the presence of a marked setting leads to a higher proportion of problem perceptions, with the exceptions of (1) FINAL MOUTH OPENING, which, when present, has a higher relative number of non-problem classifications and (2) DIVERTED HEAD POSITION, which did not have an overall influence on problem perception. If we look at the stimuli used in experiment 1 (system questions), we see that all audiovisual features have a significant influence on the perception judgments (with $p < .001$). In order of strength: FROWNING ($\chi^2 = 453.2$, $V = .437$), REPEATED HEAD GESTURES ($\chi^2 = 305.2$, $V = .358$), EYEBROW RAISING ($\chi^2 = 154.3$, $V = .255$), SMILING ($\chi^2 = 130.9$, $V = .235$), EYE MOVEMENTS ($\chi^2 = 129.2$, $V = .233$), MOUTH OPENING ($\chi^2 = 26.8$, $V = .106$) and, finally, DIVERTED HEAD POSITION ($\chi^2 = 16.7$, $V = .084$) (recall that hyperarticulation plays no role in this experiment). It is worth noting that even though DIVERTED HEAD POSITION had no overall significant effect (see Table **5**), there is a small but significant effect of this feature in the first experiment. In general, the presence of a marked audiovisual feature implies that more participants perceive problems, only for mouth opening this trend is reversed.

For the stimuli from experiment 2 (negations), the results are less clear. Only three features had a significant influence on problem perception, and in general, the scores on the Cramer's *V* test showed much weaker associations than reported for experiment 1. Ordered by strength the significant cues were: FROWNING ($\chi^2 = 43.0$, $V = .153$), HYPERARTICULATION ($\chi^2 = 31.3$, $V = .130$), and SMILING ($\chi^2 = 17.0$, $V = .096$). This outcome is consistent with the results of the perception study in experiment 2; apparently the stimuli in this part contained few cues which participants could use to determine whether a speaker's "no" came from a problematic or an unproblematic turn.

The situation for experiment 3 (destinations) is subtly different again. All features have a significant effect, apart from REPEATED HEAD GESTURES. And again, if a marked audiovisual feature setting is present, this leads to an increased proportion of perceived problems, unless the feature is MOUTH OPENING which, as above, seems to have an effect in the opposite direction. Interestingly, the relative importance of the features (in terms of strength of association) is somewhat different here: HYPERARTICULATION ($\chi^2 = 224.6$, $V = .326$), MOUTH OPENING ($\chi^2 = 87.3$, $V = .203$), FROWNING ($\chi^2 = 65.2$, $V = .176$), DIVERTED HEAD POSITION ($\chi^2 = 62.8$, $V = .172$), EYE MOVEMENTS ($\chi^2 = 7.9$, $V = .061$), EYEBROW RAISING ($\chi^2 = 6.9$, $V = .057$), SMILING ($\chi^2 = 6.5$, $V = .055$). For destinations, HYPERARTICULATION is clearly the single most important cue that participants based their perceptual judgments on.

In the presentation of the results we have treated HYPERARTICULATION as a binary cue, whereas in fact it was coded on a 7 point scale (the summed score of the 3 coders). Figure **4** shows the average proportion of participants perceiving a fragment as problematic as a function of different degrees of HYPERARTICULATION (ranging from 0 to 6), for the stimuli from experiment 2 and 3. This figure shows a clear trend, where stimuli that get more extreme values in terms of HYPERARTICULATION, also are perceived as more problematic. Correlational analysis reveals that the proportion of perceived problems increases as a function of the degree of HYPERARTICULATION (*r* = .679, *p* < .001).

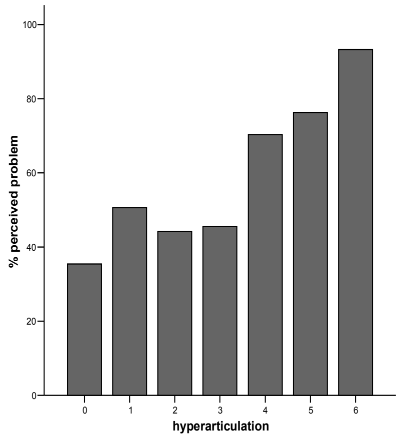

Figure **4** - Bar charts with the average proportion of participants perceiving a stimulus as problematic as a function of different degrees of hyperarticulation

In general, it appears that the presence of a marked audiovisual feature setting gives rise to more participants perceiving a problem. While the results show that there are significant effects of various features, the sizes of these effects are often rather minimal as can be seen from the Cramer's *V* scores. This suggests that the perception of problem status does not seem to be the result of a single factor in isolation. Indeed, when we checked all 2-way interactions between the various factors on the whole dataset using a multinomial logistic regression analysis, we found that all these interactions were above chance level, which suggests that perceived problem status results from a combination of cues. More detailed interaction analyses are unfortunately not feasible given the unbalanced nature of the data set and the resulting data sparseness.

As an alternative way to get a view on the effect of combinations of features, we determined if and how the perceived problem status of a stimulus depended on the number of marked features in an utterance. By focusing solely on VISUAL VARIATION, we get a better insight in the contribution of the visual factors to problem perception. To this end, we calculated the average proportion of participants perceiving a fragment as problematic as a function of the degrees of visual variation, where VISUAL VARIATION was computed by summing over the presence of marked settings of each visual feature, where SMILING, DIVERTED HEAD POSITION and REPETITIVE HEAD GESTURES were recoded in terms of presence or absence[7]. This gave a range that varied between the theoretical extremes of 0 and 7 (though we actually did not get any case where all visual features were present at the same time). The results are visualized in Figure **5**. Interestingly, the resulting picture is very similar to that in Figure **4**; more problematic fragments get more extreme values both in terms of VISUAL VARIATION and in terms of HYPERARTICULATION.

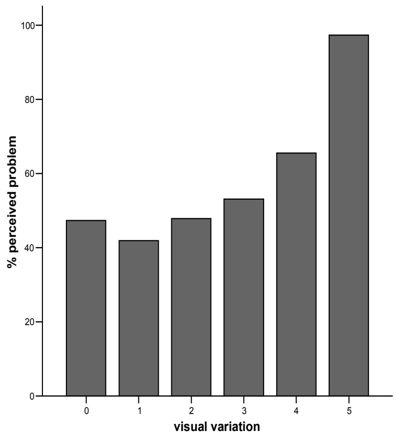

Figure **5** - Bar charts with the average proportion of participants perceiving a stimulus as problematic as a function of amount of visual variation

Audiovisual features and the presence of problems

So far we have taken a purely perceptive perspective, yet it is also interesting to take a more system-oriented perspective and investigate the relation between the audiovisual cues and the actual presence or absence of communication problems. To find out, we repeated the analysis with *problem* instead of *perceived problem* as our class of interest.

---

[7] Note that some repetitive head gestures do not appear to cue problems (e.g. nodding). In a similar vein, we saw that mouth opening is not perceived as a cue for problems either. A more sophisticated analysis to visual variation might leave out these cues, but here we simply summed over all visual variation.

The results of this analysis can be found in Table **9**, which gives the distribution of utterances from experiments 1-3 that are either problematic or not as a function of the presence or absence of a marked feature setting. The first thing to note is that we have much less data points here than in the perceptual analysis. Still, there are some significant features, namely HYPERARTICULATION ($\chi^2 = 4.8$, $V = .283$) and SMILING ($\chi^2 = 6.5$, $V = .261$).

Thus, when a speaker hyperarticulates or smiles, chances that a communication problem had occurred increase. FROWNING, EYEBROW RAISING and EYE MOVEMENTS show a similar pattern, although not statistically significant. REPEATED HEAD GESTURES and MOUTH OPENING do not seem to correlate with problem status. It is interesting to note that even though FROWNING occurs relatively often in unproblematic stimuli (12 times), participants in the perception test have a strong tendency to interpret FROWNING as a cue for problems. A somewhat similar observation can be made with respect to nodding, which occurs almost as often in unproblematic as in problematic stimuli (6 and 5 times respectively), while participants have relatively strong tendency to interpret this behavior as a cue for the absence of communication problems.

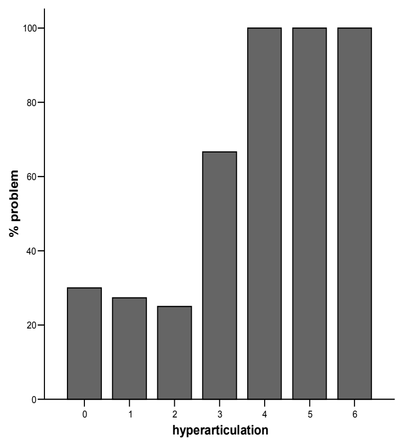

Figure **6** - Bar charts indicating the percentage of problematic stimuli as a function of different degrees of hyperarticulation

As above, it is interesting to look at both the amount of HYPERARTICULATION and at the amount of VISUAL VARIATION as cues for communication problems. Figure **6** and Figure **7** show the average proportion of problematic stimuli as a function of the amount of HYPERARTICULATION and the degrees of VISUAL VARIATION, respectively. Correlational analyses reveal that the proportion of problems increases as a function of both degree of

HYPERARTICULATION and of the amount of VISUAL VARIATION, though the latter is not significant, probably due to sparse data (HYPERARTICULATION: r = .914, p < .01; VISUAL VARIATION: r = .601, p = .207). As one would expect, HYPERARTICULATION is a clear cue for problems. But the data show a similar trend for VISUAL VARIATION: it appears to be a cue for problems as well, in the sense that if two or more visual cues are present in stimuli, the chances that the utterance was problematic increase as well. This latter bar graph also illustrates that it is not feasible to detect errors on the basis of visual cues alone, since a sizeable number of stimuli contained no visual cues but were problematic nevertheless.

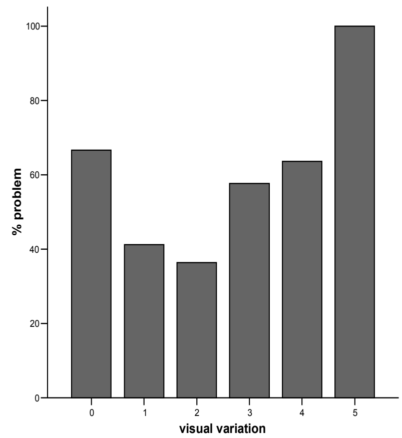

Figure **7** - Bar charts indicating the percentage of problematic stimuli as a function of the amount of visual variation

2.7.4 Summary

The main finding of the correlational analyses presented here is that the perceived problem status of a user utterance is not only reflected in a particular speech feature, i.e. in different degrees of HYPERARTICULATION, but also in the visual domain, i.e. in changes in overall facial movement. In particular, the more problematic a fragment is perceived, the more likely it has more dynamically changing auditory and visual correlates. As one would expect, there are also clear correspondences between audiovisual features and actual problem status. In particular, the *combination* of visual features is a good cue for errors. The current experiment does not allow us to determine which combinations of audiovisual features are particularly relevant for error detection, since we did not have sufficient data points to get full insight into possible interaction effects.

On the level of individual features, one interesting finding is that different features are relevant for the different experiments. For example, FROWNING and REPEATED HEAD GESTURES played a significant part in the first experiment, but had little or no effects in the third experiment. One possible explanation for this might be that in the first experiment the user listens or responds to a verification question, and thus might become *aware* of a communication problem. The stimuli in the first experiment consist of users' feedback reactions to these system verifications, and users may show surprise (FROWNING) or may (dis-)confirm the recognized information using head nodding or shaking. In the third experiment, by contrast, the users *respond* to a question from the system to provide a station name. This could be a correction, in which case HYPERARTICULATION is an important cue. This implies that a system that uses audiovisual cues for the detection of errors should look for different (combinations of) cues depending on contextual information, such as the most recent system question.

Another thing worth observing is that for nearly all individual features, the marked feature setting is associated with problems. This is perhaps surprising since many of these features are multi-interpretable. Smiling is a good example. In the current experiment, SMILING, perhaps counter intuitively, showed a positive correlation with the perception of problems. Fridlund (1994, pp.152-155) describes an experiment of Kraut and Johnston (1979), where bowlers' facial displays were analyzed after the play. The bowlers smiled more while facing friends then when looking at the pins, even when they had a bad play. This suggests that smiling can occur during a negative emotional stimulus. In the current experiment, the speaker smiled regularly (in 25 of the 96 film fragments, 26%). However, their SMILING suggested problematic interactions (17 out of the 25 fragments). A possible explanation is that there seemed to be a lot of user frustration. The smiling could have been an expression of disbelief (about the capacities of the speech recognition system). The smiling functions thus as a meta-gesture, making comments about the discourse (Kendon, 2001). In that case, the smiling might have been accompanied by other expressions as raising one's brows or frowning, resulting in a so-called blend emotion (Ekman & Friesen, 1975). As mentioned above, the feature FROWNING also had a significant correlation with the perception of problems. However, it is not clear what kind of problems the frown indicates. It is possible that it reflects the state of the discourse (the speech recognition system may just have misunderstood the speaker), but it could also reflect memory problems. It would be interesting to investigate in future studies whether the frown is the reflection of the inner state (memory overflow), or serves as a discourse signal (misunderstanding problems).

Table **9** - Distribution of utterances from experiments 1-3 that are either problematic or not as a function of the presence or absence of a marked feature setting

Feature	*Present*	*Perceived as*		*Statistics*	
		¬Problem	*Problem*	χ^2	*Cramer's V*
Hyperarticulation	No	14	6	4.8[a]	6.283
	Yes	16	24		
Smiling	No	41	30	6.5[a]	.261
	Yes	7	18		
Diverted head position	No	13	19	1.7	.133
	Yes	35	29		
Frowning	No	36	31	1.2	.113
	Yes	12	17		
Eyebrow raising	No	43	37	2.7	.168
	Yes	5	11		
Eye movements	No	26	23	.4	.063
	Yes	22	25		
Mouth opening	No	37	37	0	0
	Yes	11	11		
Repeated head gestures	No	39	40	.1	.033
	Horiz	3	3		
	Vert	6	5		

† [a] = $p < .05$. The significance and the strength of the associations are expressed in terms of $\chi 2$ ($df = 1$, except for repeated head gestures where $df = 2$) and Cramer's *V* tests, respectively

While seven of the eight features were purely labeled on a visual basis, HYPERARTICULATION was not. It would be interesting to see whether hyperarticulation can also be detected visually. It seems likely that it is indeed visible in the articulatory region. But perhaps other visual cues correlate with HYPERARTICULATION as well. It has been pointed out, for instance, that eyebrow movements are associated with accentuation (and thus perhaps with hyperarticulation as well). The current (limited amount of) data do not support this hypothesis. There are raised brows in 8 of the 40 fragments in which HYPERARTICULATION occurs (on a total of 16 raised brows), while raised brows occur in 4 of the 20 non-hyperarticulated fragments (with exclusion of 4 raised brows in study 1, as hyperarticulation was there not possible).

2.8 Discussion and conclusion

We have described three perception studies in which participants were offered film fragments (without any dialogue context) of speakers interacting with a spoken dialogue system. In half of these fragments, the speaker is or becomes aware of a communication problem. Participants had to determine by forced choice which are the problematic fragments. It was found that in all three studies, participants were capable of performing this task to a certain degree, but that the number of correct classifications varies across the three studies. As it turned out, participants had most difficulty with the second study, in which the stimuli consisted only of negation phrases ("no"). Surprisingly, the results were best in the first study, in which participants silently listen to a verification question of the system. Speculating on why the different tests have led to different results, we hypothesize that this is partly due to the fact that the stimuli in experiments 1 and 3 were longer than in experiment 2, which consisted of only a very short fragment (the word "no"). Accordingly, the longer clips may have contained more cues than the shorter ones (the mean number of marked visual cues was three for the system questions, as opposed to two in the other two studies). Next, in order to gain more insight into the audiovisual features that may have served as possible signals to problematic and unproblematic utterances and to support our preliminary informal observations, we labeled the stimuli in terms of a detailed coding scheme, comparable with (part of) the FACS system (Ekman & Friesen, 1975). It was found that, in general, each of the features had a significant effect on whether an utterance is perceived as problematic or not. The presence of a marked setting leads to a higher proportion of problem perceptions, with the exceptions of (1) FINAL MOUTH OPENING, which, when present, has a higher relative number of non-problem classifications and (2) DIVERTED HEAD POSITION, which did not have an overall influence on problem perception. In addition, *combinations* of marked feature settings are better indicators of problems than single features in isolation; more problematic fragments get more extreme values both in terms of VISUAL VARIATION and in terms of HYPERARTICULATION. Similarly, the marked feature settings also occur to a larger degree in *actual* problems, though some of the findings, due to fewer data points, represent trends rather than real significant effects.

On the basis of these results, we believe that visual information may provide a useful source for error detection, next to existing sources such as linguistic and prosodic cues. In future research, we would like to experiment with (semi-)automatic procedures to detect audiovisual cues in recordings, for instance on the basis of automatic measurements of the amount of movement and visual variation in a clip, which is potentially useful to distinguish neutral from more dynamic faces. We conjecture that such automatic facial tracking could be beneficial for improving human-machine interactions in that audiovisual correlates of problematic utterances allow systems to monitor the level of frustration of a user (Picard & Klein, 2002) or to use them as a resource for error detection.

Appendix

This appendix lists the train-travel advice tasks which were used in this chapter. Transcriptions of the one of the dialogues (with the speaker ED) associated with one of the tasks (Trajectory 3) is also given.

Tasks

Task	*Trajectory*	*Departure*
1.	Amsterdam ► Reuver	6:45 PM
2.	Utrecht ► Eijsden	6.15 PM
3.	Amsterdam ► Nijmegen	3.20 PM
4.	Nijmegen ► Swalmen	2.35 PM
5.	Utrecht ► Amsterdam	10.45 PM
6.	Tilburg ► Rotterdam	8.40 PM

Dialogue

	Dutch	*English*
System:	Vanaf welk station wilt u vertrekken?	From which station do you want to depart?
User:	Amsterdam.	Amsterdam.
System:	Is de naam van station "Amsterdam Centraal" correct?	Is the name of station "Amsterdam Centraal" correct?
User:	Ja.	Yes.
System:	Naar welk station wilt u reizen?	To which station do you want to travel?
User:	Nijmegen.	Nijmegen.
System:	Is de naam van station "Nijmegen" correct?	Is the name of station "Nijmegen" correct?
User:	Ja.	Yes.
System:	Ik heb u niet kunnen horen.	I could not hear you.
System:	Antwoord alstublieft met 'ja' of met 'nee'.	Please answer with 'yes' or with 'no'.
System:	Nogmaals: Is de naam van station "Nijmegen" correct?	Again: Is the name of station "Nijmegen" correct?
User:	Ja.	Yes.

System:	Wilt u vandaag reizen?	Do you want to travel today?
User:	Ja.	Yes.
System:	Hoe laat wilt u vertrekken?	At what time do you want to depart?
User:	Drie uur twintig.	Three o'clock twenty.
System:	Dus u wilt drie uur zeven vanmiddag vertrekken.	So you want to depart at three o'clock seven this afternoon?
User:	Nee.	No.
System:	Hoe laat wilt u vertrekken?	At what time do you want to depart?
User:	Drie uur twintig.	Three o'clock twenty.
System:	Dus u wilt drie uur twee-en-twintig vanmiddag vertrekken.	So you want to depart at three o'clock twenty-two this afternoon?
User:	Nee.	No.
System:	Hoe laat wilt u vertrekken?	At what time do you want to depart?
User:	Drie uur twintig.	Three o'clock twenty.
System:	Dus u wilt om drie uur twintig vanmiddag vertrekken.	So you want to depart at three o'clock twenty this afternoon?
User:	Ja.	Yes.
System:	De automatische reisadviseur heeft de volgende reis voor u gevonden:	The automatic travel advisor has found the following journey for you:
	U vertrekt om 15 uur 27 en komt om 17 uur 59 aan. De reistijd is 1 uur 32 met 1 overstap.	You depart at 15 o'clock 27 and arrive at 17 o'clock 59. The traveling time is 1 hour 32 with 1 change.
System:	Wilt u het reisadvies gedetailleerd horen?	Do you want to hear the detailed travel advice?
User:	Nee.	No.

3 End-of-utterance detection

in the auditory and the visual modality

3.1 Introduction

The last chapter described how communication problems are reflected in audiovisual speech during a human-machine dialogue. The current chapter discusses how audiovisual speech is used to signal the end of an utterance, which may play a role in the fluency of turn-taking.

Speakers use non-lexical features to demarcate various kinds of speech units, varying from a simple phrase to a larger scale discourse segment or a turn in a natural conversation[8]. Previous studies have largely focused on how prosodic variables, such as intonation, rhythm and pause, or more subtle modulations of voice quality, like creaky voice, can be exploited to signal the end of such units (e.g. de Pijper & Sanderman, 1994; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Swerts, Bouwhuis, & Collier, 1994; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). In addition to features that are encoded in the speech signal itself, there is also an investigation into how particular visually observable variations from a speaker's face, like gaze patterns or bodily gestures, can be used as boundary cues (e.g. Argyle & Cook, 1976; Cassell, Nakano, Bickmore, Sidner, & Rich, 2001; Nakano, Reinstein, Stocky, & Cassell, 2003; Vertegaal, Slagter, van der Veer, & Nijholt, 2000). However, little is known about the perception of these visual cues, and about the relative importance of the visual and the auditory modality for demarcation purposes. Therefore, the aim of this chapter is to get more insight into which modalities speakers use for signaling finality or non-finality, and how sensitive observers are to these respective signals. In particular, our goal is to investigate the relative contribution of three different conditions to end-of-utterance detection: two unimodal ones, vision-only and audio-only, and their bimodal combination.

---

[8] An earlier version of this chapter was published in Barkhuysen, P., Krahmer, E., & Swerts, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America*, *123*(1), 354-365.

It is by now well-established that various auditory cues may serve as boundary markers of speech utterances (e.g. Koiso, Horiucho, Tutiya, Ichikawa, & Den, 1998; de Pijper & Sanderman, 1994; Swerts, Bouwhuis et al., 1994; Ward & Tsukahara, 2000; Wightman et al., 1992, among many others). One of the strongest prosodic indicators for the end of a speaker's utterance is a pause, either a silent interval or a filler such as "uh" and "uhm", (as shown by, among others, de Pijper & Sanderman, 1994; Price et al., 1991; Swerts, 1997, 1998; Wightman et al., 1992). Many of these studies are based on analyses of monologues, where it was even found that pause length may co-vary with the strength of a boundary. When looking at natural interactions between multiple speakers, however, pauses tend to be rather short in between two consecutive speaker turns. Even though end-of-utterance pauses may be very short in interaction, turn switching proceeds remarkably smoothly, generally without overlap between speakers (Koiso et al., 1998; Levinson, 1983, pp.296-297; Ward & Tsukahara, 2000).

One of the reasons why the turn-taking mechanism may proceed so fluently, is that speakers "presignal" the end of their utterances (e.g. Caspers, 1998; Couper-Kuhlen, 1993; Swerts, Bouwhuis et al., 1994; Swerts, Collier, & Terken, 1994). Listeners may pick up these cues and therefore may know in time when the current turn will be finished. Various researchers have looked in detail at the nature of these cues. It has been suggested, for instance, that the capacity of listeners to feel an upcoming boundary is based on what is called rhythmic expectancy, which would steer turn-taking to some extent (Couper-Kuhlen, 1993). Related to this, there is subtle durational variation, such as preboundary lengthening, which speakers can use to mark the final edge of a speech unit such as a turn (Price et al., 1991; Wightman et al., 1992). In addition to these timing-related phenomena, many researchers have focused on the potential use of melodic boundary markers as well. First, there are local boundary markers which occur at the extreme edge of a turn-unit, right before an upcoming boundary, for which it has been shown that tones which reach a speaker's bottom range clearly function as finality cues (Caspers, 1998; Koiso et al., 1998; Swerts & Geluykens, 1994). Moreover there appear to exist melodic structuring devices which are more global in nature in that they are spread over a whole speech unit. In particular, various studies have pointed out that speech melody gradually decreases in the course of an utterance, which may enable listeners to feel a boundary coming up (e.g. Leroy, 1984). However, this declination pattern may be typical of read-aloud speech which allows for a larger degree of look-ahead compared to spontaneous speech. Other finality cues are variations in pitch span, and more subtle differences in the alignment of pitch movements (Silverman & Pierrehumbert, 1990; Swerts, 1997). Finally, there is acoustic evidence which shows that marked deviations from normal phonation, in particular, creaky voice, typically occur at the end of an utterance (Carlson et al., 2005).

The possible pre-monitoring cue value of prosodic cues has been explicitly tested in various perception studies. Grosjean (1983) and Leroy (1984) have already established that human participants are surprisingly accurate in estimating the location of an upcoming boundary, using a variant of a gating paradigm, in which listeners are only presented with the initial part of an utterance. Along the same lines, Swerts, Bouwhuis et al. (1994) and Swerts & Geluykens (1994) reported that people are able, on the basis of melodic cues, to judge the serial position of a phrase in a larger discourse unit. Carlson et al. (2005) found that native speakers of Swedish and of American English showed a remarkable similarity in judgments when they had to predict upcoming prosodic breaks in spontaneous Swedish speech, even when they had to base such estimations on stimuli which consisted of only a single word.

It thus seems safe to conclude that speakers and listeners take the auditory modality into account while marking the end of an utterance. But to what extent do they pay attention to the visual modality? Various researchers have argued that speakers may use visual cues for end-of-utterance signaling, where most studies have investigated how various bodily gestures may be used as markers of discourse boundaries. First, different studies focused on general changes in posture (Argyle & Cook, 1976, p.101; Cassell et al., 2001; Duncan, 1972). These studies suggest that there is a general trend for people to change their pose when they start speaking, whereas they return to their initial posture at the end of a turn, for instance by raising their shoulders at the onset of a turn and lowering them again at the end, or by changing their head orientation (McClave, 2000). Second, one specific visual cue which has received much scholarly attention is related to movements of the eyes. Argyle & Cook (1976, pp.114-120) describe in detail how the tuning of gaze behavior regulates many aspects of the interaction in a very subtle way. In general, it appears to be the case that speakers divert their gaze rather often while talking, whereas the listening conversation participant tends to look at the partner more frequently. When analyzing the gaze patterns in normal interactions more closely, it appears that a pattern emerges which is connected to the turn-taking mechanism, in that speakers tend to divert their gaze when they start talking, and return the gaze to their partner when they are finished (see also Beattie, Cutler, & Pearson, 1982; Goodwin, 1980; Kendon, 1967; Nakano et al., 2003; Novick, Hansen, & Ward, 1996; Vertegaal et al., 2000). The cue value of gaze is likely to be due to the fact that human eyes have a unique morphology, with a large white sclera surrounding the dark iris. It has been argued that this contrast may have evolved to make it easier to detect the gaze direction of other's (Kobayashi & Kohshima, 1997). While variation in posture shifts and gaze patterns have been directly linked to boundary marking, in particular in the turn-taking system, various researchers have argued that there may be further visual cues which may be important for demarcation purposes as well, such as head nods (e.g. Maynard, 1987), eyebrow movements (e.g. Cavé et al., 1996; Ekman, 1979; Krahmer & Swerts, 2004), and eye blinks (e.g. Doughty, 2001).

The results from the various studies described above thus suggest that a speaker can display that (s)he is going to stop speaking, by means of both auditory and visual features. However, there are still a large number of unsolved questions regarding the relative importance of the modalities and of their combined effects. While it has been shown that listeners are accurate in determining the end of an utterance based on the auditory modality, it is unknown whether they would be equally capable to do so at the basis of visual information. And if so, it is still an empirical question as to how the visual modality relates to the auditory one, whether or not the two modalities may reinforce each other, and whether observers are helped or rather distracted when they have to focus on two rather than on a single modality in their finality judgments.

To this end, we have set up two experiments that are both based on perceptual judgments of stimuli in one of three conditions: a VISION-ONLY, AUDIO-ONLY or an AUDIOVISUAL condition. The experiments make use of audiovisual recordings of semi-spontaneous utterances that were naturally elicited in a question-answering paradigm. The first experiment explores differences between modalities via a reaction time experiment in which participants are instructed to indicate as soon as possible when they think an utterance, presented in one of three conditions, ended. The second experiment makes use of basically the same stimuli as the ones from the first experiment, and looks in more detail at which factors influence participants' abilities to judge whether a speaker's turn is about to end or not; in this experiment, participants are presented both with longer and shorter speech fragments, so we may get insight into the cue value of possible global versus local cues to finality. In addition, we look in more detail into the question of which auditory and visual cues are actually used by our speakers.

3.2 Audiovisual recordings

We gathered digital video recordings of speakers responding to questions in a natural, interview-style situation. Although recent research suggests that lexical and syntactic factors are relevant for end-of-utterance detection (de Ruiter et al., 2006), for our current purposes, however, these factors should be eliminated as they would offer an unfair advantage to the auditory modality. Hence the questions were intended to elicit lists of words, where the lexical and syntactic structures of the answers offer no clues at all about where the end of the utterance is to be expected.

The questions were selected in such a way that they resulted in a variety of different answers, and such that potential answer words could occur in different positions in the list, depending on the question. Target answers varied in length, consisting of three or five words. Twelve questions were asked for predictable sets of numbers, in different orders and with different number ranges. For instance,

- What are the multiples of five below thirty?
- What are the odd numbers below ten in reversed order?
- What are the multiples of five below thirty in reversed order?

Notice that the word "five" can occur both in a FINAL and in a NON-FINAL position. The other questions addressed general knowledge or individual preferences of the interviewee, such as:

- What are the colors of the Dutch flag?
- What are your three favorite colors?
- Name five countries where you can go skiing

Notice that for the second category the answers are never fully predictable. Even the colors of the Dutch flag are described by participants both as "red, white, blue" and "blue, white, red". Moreover, both "red" and "blue" can occur (and do in fact occur) as the second, middle, word, in responses to the favorite color question. The interview consisted of 33 questions, of which 25 were experimental and 8 were filler items. As filler items, questions were used for which the number of words in the answers could in principle not be predicted (e.g. "Which languages do you speak?"). These filler items were added for the sake of variety and to make sure that speakers did not only produce three and five word lists.

A total of 22 speakers participated (13 male and 9 female), between 21 and 51 years old. None of the speakers was involved with audiovisual research, and speakers did not know for what purpose the data were collected. The original recordings were made with a digital video camera (MiniDV; 25 frames/s, a resolution of 720 × 576 pixels, sampling of 4:2:0 (PAL), luma 8 bits chroma and 2 channel audio recording at 16 bits resolution and 48 kHz sampling rate). The recordings were subsequently read into a computer and orthographically transcribed. See Figure **8** for some representative stills.

3.3 Experiment 1: Reaction times

As a first exploration we performed a reaction time experiment with the intention to gain insight into the relative contribution of the auditory and visual modality, alone and in combination, for end-of-utterance detection.

3.3.1 Stimuli

For this experiment 4 male and 4 female speakers were randomly selected from the corpus of 22 speakers described above. For each speaker, three instances of answers consisting of 3 words and three instances of 5 words were randomly selected on the basis of the transcriptions (8 speakers × 6 instances = 48 stimuli in total). Notice that since this first selection was random, the set of selected answers differed for each of the selected speakers. As a result, the lexical content of the selected answer lists was highly varied, and since words could occur in various (FINAL and NON-FINAL) positions, participants could never rely on lexical information for their end-of-utterance detection.

If the first selection contained answers with more than just list words (e.g. repetitions of the question, or fragments where speakers think aloud), these were replaced with another randomly selected answer. Moreover, lists where the pre-final and final word were separated by a conjunction (i.e. lists of the form "A, B and C") were replaced as well. In addition, for each speaker two filler items were selected of different lengths. Fillers could include other spoken text (such as repetitions or corrections), and as a result the average length of filler items is 11 words. Each stimulus was cut from the interview session in such a way that it started immediately after the interviewer finished asking the current question until 1000 *ms* after the speaker finished answering (i.e. 1000 *ms* after the auditory speech signal of the answerer had stopped).

Figure **8** - Representative stills of speakers SS (top) and BB (bottom) while uttering the first and middle word and just after uttering the final word of a three word answer, such as "red, white, blue."

3.3.2 Design

The experiment had a counterbalanced 3 × 3 Latin square within-subjects design, with condition (with 3 levels: one bimodal, containing AUDIOVISUAL stimuli (AV), and two unimodal ones, one AUDIO-ONLY (AO) and one VISION-ONLY (VO)) and stimulus duration (with levels: 3-WORD and 5-WORD) as within-subjects factors, and reaction time as the dependent variable.

3.3.3 Procedure

Stimuli were presented to participants in three conditions: one bimodal one, containing audiovisual stimuli (AV), and two unimodal ones, one audio-only (AO), and one vision-only (VO). In the AUDIOVISUAL condition, participants saw the stimuli as they were recorded. In the AUDIO-ONLY condition, participants heard the speakers while the visual channel only depicted a static black screen, and in the VISION-ONLY condition, participants only saw the speakers but could not hear them. All participants entered all three conditions (within-design), but the order in which participants entered these conditions was systematically varied (using a 3 x 3 Latin square design). Moreover, within a condition, stimuli were always presented in a different random order. In this way, all potential learning effects could be compensated for.

Each condition consisted of two parts: a baseline measurement and the actual end-of-utterance detection. Each part was preceded by a short practice session to make participants acquainted with the experimental setting and the kind of stimuli in the current condition. The practice session did not contain lexical material which reoccurred in the actual experiment.

The aim of the baseline measurement was to find out how long it took participants on average to respond to comparable stimuli in the three modalities of interest (AV, AO, VO) of varying durations but always completely devoid of finality cues. During the baseline measurement, the participants' task was to press a designated button as soon as the end of the stimulus was reached. Stimuli were constructed to make them comparable to the actual stimuli used in the non-baseline conditions but without introducing potential finality cues. In the AUDIOVISUAL modality, the baseline stimuli therefore consisted of a video still (a single frame of some speakers[9]) accompanied by a stationary /m/ (a male voice for male speakers, and a female voice for female speakers), creating the impression of a speaker uttering a prolonged "mmm". In the VISION-ONLY baseline measurement, only the video still was displayed, and in the AUDIO-ONLY baseline measurement, only the stationary /m/ was heard. In all three conditions the baseline stimuli are therefore completely static: the face does not move, since it is a still image, and the sound does not change either, since it is stationary. When the end of a baseline stimulus is reached, the sound stops (in the AO condition) and a blank screen appears (in the VO condition); this happens simultaneously in the AV condition. Only then can participants know that the stimulus ended; there is no conceivable cue in the stimulus which could presignal this.

During the actual end-of-utterance detection part, participants were instructed to indicate, as soon as possible, when the speaker finished his or her utterance by pressing a dedicated button. In the experiment, it was crucial that participants pay attention to visual information on the screen. Therefore, they were given an additional monitoring task, where participants had to press another button as soon as they saw a small red dot appearing on the screen. These red dots were added to a limited number of dummy stimuli. Even though the AUDIO-ONLY condition did not include any potentially relevant visual information (only a black screen), participants also had to spot the red dots in this condition to make sure all conditions were alike in this respect. The duration of the red dot appearance was 1/25 *s* (a single frame); it appeared at varying locations on the screen. The dummy stimuli were only used to control the visual attention of participants and were not used in the reaction time

[9] We explored the use of dynamical visual material containing non-speech sounds, such as laughter, but this did not seem suitable, because the perceivers were distracted by the emotional content of that material, and more important, the non-speech material did not seem to be devoid of finality cues.

analyses. This use of dots to make sure that the participants will process the visual information is a common procedure in audiovisual speech research (e.g. Bertelson, Vroomen, & de Gelder, 2003).

The experiment was individually performed. Participants were invited into a quiet room, and asked to take a seat behind a computer on which the stimuli would be displayed. There were loudspeakers to the left and right of the screen through which the sound was played. Participants received instructions before each of the three conditions and before they started with the relevant practice session. If everything was clear, the actual experiment started and the experimenter moved out of the visual field of the participant. There was no further interaction between participant and experimenter during the experiment.

3.3.4 Participants

For the reaction time experiment, 30 right-handed native speakers of Dutch participated, 7 male and 23 female, between 24 and 62 years old. None of the participants had participated as a speaker in the data collection phase, and none was involved in audiovisual speech research.

3.3.5 Statistical analyses

Reaction times (RT's) were always measured in milliseconds from the actual end-of-utterance (i.e. the moment where the speech signal ended). An RT of 0 thus means that a participant pressed exactly at the end of the utterance (when the auditory speech signal stopped). Notice that in the baseline measurement, the end of the dummy utterance /mmm/ also marked the end of the stimulus. In the actual experiment, stimuli continued for 1000 *ms* after the speaker finished speaking (i.e. after the spoken audio signal ended), and the end-of-utterance thus does not coincide with the end of the stimulus[10].

Inspection of the measurements revealed that occasionally a negative RT was recorded. This happened 13 times during the baseline measurement (i.e. 1.8% of the baseline data

[10] This is because we assumed that it could be possible that the visual end-of-utterance may not be in exact synchrony with the auditory part, but that the speaker's facial expressions return to their rest position either before or after the end of the auditory speech signal. It has been shown in monkeys, for example, that the onset of a facial expression (such as lip movements) starts before the onset of the acoustical signal (Ghazanfar et al., 2005). Also, left and right parts of the face do not have the same timing (Hauser & Akre, 2001).

points), and 302 times during the actual experiment (nearly 7% of the experimental data points). In both cases, the negative RT's were evenly distributed over the modality conditions. In the case of the baseline measurement we can be certain that these are errors, since participants had to respond to the "ending" of the baseline stimuli and, as explained above, there were no cues that could possibly pre-signal the end[11]. Hence these errors were replaced by the mean RT value for that stimulus. It is important to note that this did not significantly alter the results, so the inclusion of the negative RT's in the baseline condition would have led to basically the same results as reported below (given the very small number of negative instances).

In the actual end-of-utterance experiment a negative RT is not necessarily an error, because here, as noted in the introduction of this chapter, pre-signals may occur, and hence the participant may feel the end of the utterance is near even though the speaker has not actually stopped speaking yet. Since there is no other criterion for their exclusion, we decided not to remove these negative RT's. Finally, there was a total of 23 non-responses (0.5%), which were treated as missing values in the statistical analysis. We did not manipulate the raw data in any other way.

3.3.6 Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse-Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. *Post hoc* analyses were performed with the Bonferroni method.

3.3.7 Results

A general overview of the RT results for the different conditions can be found in Table **10**. First consider the *baseline* measurement. Here the VISION-ONLY (VO) condition evoked the fastest reaction times followed by the AUDIO-ONLY (AO) and the AUDIOVISUAL (AV) conditions.

---

[11] Because - as opposed to in the actual experiment - the signal was static: containing a video still and a monotonous sound.

Table **10** - Reaction times in milliseconds for the different conditions (AV, audiovisual; VO, vision-only; AO, audio-only) in both the baseline measurement and the actual experiment, with standard errors and with 95% confidence intervals

Measurement	*Condition*	*RT*	*Std. error*	*95% CI*
Baseline	AV	391.7	7.6	(376.1, 407.3)
	VO	330.8	5.9	(318.9, 342.9)
	AO	380.3	5.5	(368.9, 391.7)
Experiment	AV	508.8	38.6	(429.7, 587.8)
	VO	668.5	33.3	(600.4, 736.7)
	AO	524.6	40.2	(442.4, 606.9)

An ANOVA with condition and stimulus duration as within-subjects factors and reaction time as the dependent variable was performed. It indeed revealed a main effect of *condition* (F(2, 58) = 11.215, p < .001, η^2_p = .279). *Post hoc* analyses showed that there was a significant difference between the AUDIOVISUAL and VISION-ONLY condition (p < .001), and between the VISION-ONLY and the AUDIO-ONLY condition (p < .001). The AUDIO-ONLY and the AUDIOVISUAL condition did not, however, differ significantly (p = .368). The stimuli in the baseline variant differed in *duration*, but this did not have a significant influence on the reaction times (F(7, 203) = 2.891, n.s.), nor was the interaction between *condition* and *stimulus duration* significant (F(14, 406) = 2.021, n.s.).

Table **11** - Reaction times in milliseconds for the different conditions (left column: AV, audiovisual; VO, vision-only; AO, audio-only) in the actual experiment as a function of length (top row: 3 words or 5 words), with standard errors between brackets

Condition	*Three words*	*Five words*
AV	585.0 (36.6)	432.5 (42.7)
VO	803.9 (33.0)	533.0 (44.3)
AO	627.6 (48.9)	421.7 (42.6)

Next consider the results of the actual *experiment*. Here the AUDIOVISUAL (AV) condition yielded the quickest responses, followed by the AUDIO-ONLY (AO) condition, while the VISION-ONLY (VO) condition leads to the slowest reaction times. An ANOVA with condition, length (measured by the number of words: three or five), and speaker as within-subjects factors and reaction time as the dependent variable was carried out. A significant main effect of *condition* was found (F(2, 58) = 17.052, p < .001, η^2_p = .370). *Post hoc* analyses showed that there was a significant difference between the AUDIOVISUAL and VISION-ONLY condition ($p < .001$), and between the VISION-ONLY and the AUDIO-ONLY condition ($p < .001$). The AUDIO-ONLY and the AUDIOVISUAL condition did not differ significantly ($p = .396$). In addition, a main effect of *stimulus length* was found ($F(1, 29) = 90.086$, $p < .001$, $\eta^2_p = .756$). Inspection of Table **11** reveals that 3 WORD utterances led to longer reaction times than 5 WORD utterances. Finally, there was also a main effect of *speaker* ($F(7, 203) = 23.500$, $p < .001$, $\eta^2_p = .448$) which indicates that some speakers gave overall better or more cues that they were approaching the end of the utterance than other speakers did.

When looking at the interaction effects, a significant interaction between *condition* and *stimulus length* ($F(2, 58) = 26.480$, $p < .001$, $\eta^2_p = .477$) was found. As can be seen in Table **11**, the RT for 3 WORD utterances and for 5 WORD utterances differs substantially across the different conditions: it is relatively small for the AUDIOVISUAL condition and relatively large for the VISION-ONLY condition, suggesting that the presence of extra cues in longer fragments is particularly useful for the VISION-ONLY condition.

The RT patterns for the eight speakers are similar over the three modality conditions, as can be seen in Figure **9**. However, some speakers score particularly well in one of the conditions, for instance, because they better cue the end of their utterances using facial cues rather than auditory ones.

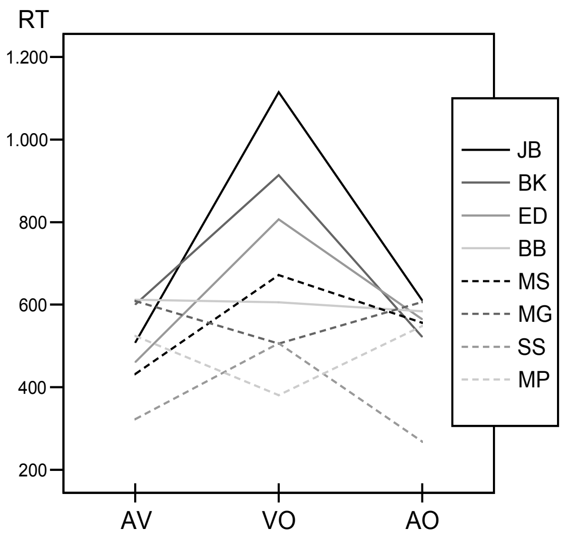

Figure **9** - The mean reaction time (ms) for the different speakers in the three modalities

It is interesting to see that the reaction time patterns for the baseline measurement are rather different from those of the actual experiment. The aim of the baseline measurement was to find out how long it takes to respond to a stimulus without any finality cues presented in a certain modality, and to compare these scores with the reaction times in the actual experiment, in order to eliminate the influence of the presentation modality itself. The picture that emerges is visualized in Figure **10**, which shows that the reaction times for the baseline and non-baseline versions are more similar in the AUDIOVISUAL condition, and more divergent in the VISION-ONLY condition, while the results for the AUDIO-ONLY condition are in between these two extremes. That is, where the visual modality leads to the fastest RT results in the baseline measurement, they are the slowest in the actual experiment. The reverse is true for the data in the audiovisual modality, whereas the data for the auditory modality are in the middle in both sessions.

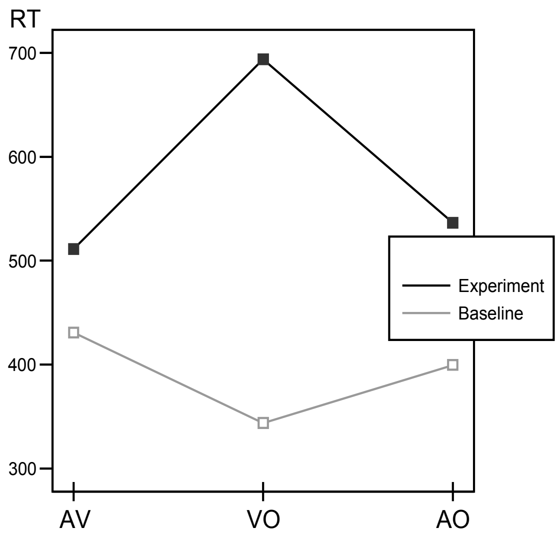

Figure **10** - The mean reaction time (ms) in the three conditions for the baseline and the actual experiment

To test these differences, we computed a difference score for each participant and stimulus, by subtracting the AUDIOVISUAL baseline RT scores for that participant from his or her non-baseline RT scores for the AUDIOVISUAL stimuli, and similar for stimuli in the other two modalities. The resulting average difference score was 80.3 *ms* for the AUDIOVISUAL condition, 136.8 *ms* for the AUDIO-ONLY condition and 349.9 *ms* for the VISION-ONLY condition. We then performed a univariate ANOVA with *average difference score* for each participant as dependent variable, and *condition* (AV, AO and VO) as independent variable, which indeed revealed a significant effect of *condition* on difference score ($F(2, 87) = 13.704$, $p < .001$, $\eta^2_p = .40$). A Bonferroni *post hoc* analysis revealed that all pairwise comparisons were significant at the $p < .001$ level, except the one between the AUDIOVISUAL and the AUDIO-ONLY condition ($p = .906$).

3.3.8 Summary

In the first experiment, we measured reaction times for end-of-utterance detection in three different conditions: AUDIO-ONLY, VISION-ONLY and AUDIOVISUAL. If prediction of the end of a turn was impossible, the reaction times for the different modalities in the actual experiment would have been the same, or at least have the same pattern as in the baseline measurement, where no cues were present. However, this is clearly not what was found. Rather, the AUDIOVISUAL stimuli in the actual experiment led to the quickest responses, the AUDIO-ONLY stimuli led to slightly longer reaction times (although the difference with the AUDIOVISUAL stimuli was not statistically significant), and the VISION-ONLY stimuli led to the slowest responses. While this result suggests that combining modalities is useful for end-of-utterance detection, it also leaves open the possibility that participants essentially rely on auditory information only for end-of-utterance detection. This issue is investigated more closely in a second experiment, where participants have to classify brief fragments as non-final or final (end-of-utterance) ones.

3.4 Experiment 2: Classification

The design of the classification task resembles the design used in gating tasks. In a gating task a spoken language stimulus is presented in segments which increase in length, usually starting at the beginning of the stimulus. Participants must try to recognize the entire spoken stimulus on the basis of the fragment (Grosjean, 1996). In one possible presentation format, the *duration-blocked format*, participants hear all the stimuli at a particular segment size, then all the stimuli again in a different segment size (Grosjean, 1996; Walley, Michela, & Wood, 1995). In the current experiment we used two sizes, a LONG and a SHORT one, both of which did not cover the entire original utterance. Participants had to make a binary decision about the setting from which the fragment originated (i.e. FINAL or NON-FINAL).

3.4.1 Stimuli

The stimuli for experiment 2 were selected from the utterances of the same 8 speakers which were used in experiment 1. For each of these speakers we randomly extracted answers from their original set of answers (see section 3.2), and constructed two types of fragments from these: LONG ones, consisting of 2 words, and SHORT ones, consisting of 1 word. For each of the eight speakers, we created 4 LONG pairs (final/non-final) and 4 SHORT pairs of fragments, where the SHORT fragments always consisted of the last word of the corresponding LONG (2 word) fragment[12].

Orthogonal to this, half of the fragments were from a FINAL (end-of-utterance) and half from a NON-FINAL position. In the same way as for experiment 1, we made sure that participants could not pick up on lexical cues for their final/non-final classifications. Naturally, the FINAL pairs were always selected from the tail of the list, while the NON-FINAL pairs were selected from varying positions in the list. The length of the original context surrounding a fragment was more or less balanced, with a small majority of fragments extracted from answers consisting of five words.

To guarantee the understandability of the fragments and to make sure they were comparable across conditions, the fragments were selected such that they included a naturally occurring pause after the last word of the fragment (when it was a NON-FINAL fragment), or a pause after the end of the original answer (when it consisted of the FINAL part of an answer). The fragments were always cut in such a way that the pauses in the NON-FINAL fragment and the corresponding FINAL fragment in the corresponding 1 word (SHORT) and 2 word (LONG) stimuli lasted equally long, to make sure that the length of the pause (which, as noted in the introduction of this chapter, is an important signal for end-of-utterance) could not be used as a cue for classification. Also, the pauses in the NON-FINAL fragment and the corresponding FINAL fragment were equally long, when they were derived from the same original answer.

3.4.2 Design

As for experiment 1, all fragments were stored in three ways: AUDIO-ONLY (AO), VISION-ONLY (VO) or AUDIOVISUAL (AV). Therefore, in total 128 stimuli were created for each modality: 8 speakers × 2 lengths (SHORT and LONG) × 2 types (NON-FINAL and FINAL) × 4 fragments.

[12] In other words: for each long fragment, a short fragment was cut from the last half of the long (2 word) fragment, and these long fragments were selected from non-final as well as final positions.

Again, stimuli were presented in three conditions: an AUDIOVISUAL (AV), an AUDIO-ONLY (AO) and a VISION-ONLY (VO), which were presented to participants in the same format as in experiment 1, but this time in a between-subjects design.

3.4.3 Procedure

Participants were given a simple classification task: they were told to determine for each fragment whether it marked the end of a speaker's utterance or not. Each condition consisted of two parts: one part for the SHORT (1 word) fragments and one part for the LONG (2 word) fragments. The order in which participants passed the two different parts was systematically varied. For each part, two lists were created with a different random order. Participants were exposed to either the A-versions or the B-versions of a list. Therefore, each participant passed the items in a different random order in each part, and, due to the Latin square design, since the order in which participants underwent the short and long fragments part was also systematically varied, potential learning effects could be compensated for.

Each condition was preceded by a short practice session, consisting of two stimuli (different from the experimental stimuli), so that participants could get used to the type of tasks and stimuli. The general procedure was the same as for experiment 1.

3.4.4 Participants

The participants consisted of a group of 60 native speakers of Dutch: 25 male and 35 female, between 20 and 56 years old. None of them participated as a speaker in the data collection phase nor as a participant in experiment 1, and none was involved in audiovisual speech research.

3.4.5 Statistical analyses

Tests for significance were performed with a repeated measures analysis of variance (ANOVA), with speaker (eight levels), stimulus length (SHORT: 1 word, LONG: 2 words), and fragment type (NON-FINAL and FINAL) as within-subjects factors and modality (VISION-ONLY: VO, AUDIO-ONLY: AO, and AUDIOVISUAL: AV) as a between-subjects factor (mixed design) and with the percentage of correct classifications over the four fragments as the dependent variable (recall that for each speaker four short and long pairs of FINAL and NON-FINAL stimuli were selected). Mauchly's test for sphericity was used to test for homogeneity of variance, and when this test was significant or could not be computed, we applied the Greenhouse-

Geisser correction on the degrees of freedom. For the purpose of readability, we report the normal degrees of freedom in these cases. The Bonferroni correction was applied for multiple pairwise *post hoc* comparisons, and contrasts were computed in several cases.

3.4.6 Results

Table **12** gives the overall results for three factors of interest, i.e. modality, fragment type and stimulus length. According to the ANOVA, all three factors had a significant influence on the classification.

Table **12** - For each factor, the levels of the factor, the percentage of correct judged utterances with standard errors and 95% confidence intervals are given

Factor	*Level*	*% correct*	*Std. Error*	*95% CI*
Modality	AV	84.7	0.11	(82.5, 86.9)
	VO	75.7	0.11	(73.6, 77.9)
	AO	73.6	0.11	(71.5, 75.8)
Fragment type	NF	80,8	0.11	(78.6, 83.0)
	F	75.2	0.12	(72.9, 77.7)
Stimulus length	Short	75.1	0.09	(73.3, 77.0)
	Long	81.0	0.07	(79.5, 82.3)

The most interesting main effect is that of *modality*, which was significant as well ($F(2, 57) = 29.475$, $p < .001$, $\eta^2_p = .508$). It is interesting to note that both unimodal conditions yield around 75% correct classifications (75.7 for the VISION-ONLY condition and 73.6 for the AUDIO-ONLY condition), and that both are clearly outperformed by the bimodal, AUDIOVISUAL condition (with 84.7% correct).

Post hoc analyses showed that there was a significant difference between the AUDIOVISUAL and the VISION-ONLY condition ($p < .001$), and between the AUDIOVISUAL and the AUDIO-ONLY condition ($p < .001$). The VISION-ONLY and the AUDIO-ONLY condition did not, however, differ significantly ($p = .54$). This pattern of results is visualized in Figure **11**.

Next, consider the main effect of *fragment type* ($F(1, 57) = 7.855$, $p < .01$, $\eta^2_p = .121$). It appears that judging non-finality is somewhat easier than judging finality (80.8% vs. 75.2%), but overall it is clear that the vast majority of the fragments is classified correctly. *Stimulus length* also had a significant influence ($F(1, 57) = 28.800$, $p < .001$, $\eta^2_p = .336$). Inspection of

Table **12** reveals that SHORT (1 word) fragments are somewhat more difficult than LONG (2 word) fragments.

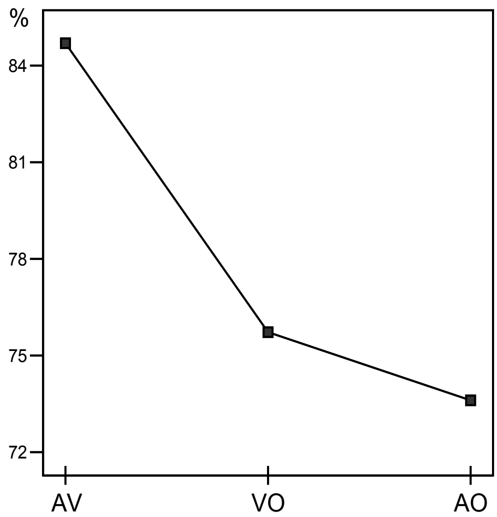

Figure **11** - Percentage of correct answers in the audiovisual (AV), vision-only (VO) and audio-only (AO) conditions

Besides the main effects for the three factors listed in Table **12**, the factor *speaker* also had a significant main effect ($F(7, 399) = 52.375$, $p < .001$, $\eta^2_p = .48$). As can be seen in Table **13**, the total number of correct classifications differs per speaker, ranging from 63% correct for speaker JB to 87.8% for speaker SS. *Post hoc* analyses showed that this difference was significant ($p < .001$). Various other pairwise comparisons of speakers were significant as well, and this shows that there are overall substantial differences between speakers in end-of-utterance signaling. It is rather interesting to observe that the scores per speaker may differ across conditions. Indeed, a significant 2-way interaction was found between *speaker* and *modality* ($F(7, 399) = 14.764$, $p < .001$, $\eta^2_p = .341$); in Table **13** it can be seen that, for instance, speaker BB apparently offers clearer visual than auditory cues, as the percentage of correctly classified stimuli for this speaker drops considerably in the AO condition. This is different for speaker MG, for instance, who seems to send more useful auditory cues (in her case the classification scores drop in the VO condition). Simple contrasts showed that this difference was significant ($F(2, 57) = 78.839$, $p < .001$, $\eta^2_p = .734$).

Table **13** - For each speaker, the total percentage of correctly judged utterances, and the percentage of correctly judged utterances as a function of the 3 modalities

Speaker	*AV*	*VO*	*AO*	*Total*
BB	86.5	86.5	56.8	76.7
BK	74.1	74.4	59.3	69.3
ED	90.6	73.3	77.7	80.5
JB	64.7	57.5	66.9	63.0
MG	86.6	68.1	86.0	80.2
MP	85.9	76.7	76.2	79.6
MS	93.1	87.2	81.0	87.1
SS	96.2	82.0	85.0	87.8

In addition, a significant two-way interaction was found between *fragment type* and *stimulus length* ($F(1, 57) = 11.317$, $p < .01$, $\eta^2_p = .166$). This interaction can also be explained by looking at Table **14**, where it can be seen that for the NON-FINAL fragments, the LONG stimuli evoked more correct answers (85.7%) than the SHORT stimuli (75.9%), while for the FINAL fragments the stimulus length makes almost no difference (74.3% versus 76.2% resp.).

Table **14** also illustrates a second, significant 2-way interaction, between *stimulus length* and *modality* ($F(2, 57) = 6.889$, $p < .01$, $\eta^2_p = .195$). As expected, for both stimulus lengths, the AUDIOVISUAL modality is the easiest one. For the SHORT fragments, the AUDIOVISUAL modality (82.5% correct answers) is followed by the VISION-ONLY modality (74.9%), and subsequently the AUDIO-ONLY modality (67.9%). A *post hoc* test within the SHORT fragments revealed that all pairwise comparisons are statistically significant (AV-VO, $p < .01$, AV-AO, $p < .001$, and VO-AO, $p < .05$). However, for the LONG fragments, the AUDIOVISUAL modality (86.9% correct answers) is followed by the AUDIO-ONLY modality (79.4%), and subsequently the VISION-ONLY modality (76.6%). A *post hoc* test within the long fragments revealed that all pairwise comparisons differ at the $p < .001$ level, with the exception of the difference between VO and AO which is not significant. No other significant interactions were found.

Table **14** - For each modality, the percentage of correctly judged utterances, as a function of stimulus length (1 or 2 words) and fragment type (non-final and final)

Length	*Finality*	*AV*	*VO*	*AO*	*Total*
1	NF	81.8	76.2	69.7	75.9
	F	83.1	73.6	66.0	74.3
Subtotal		82.5	74.9	67.9	
2	NF	89.4	82.6	85.2	85.7
	F	84.5	70.6	73.6	76.2
Subtotal		86.9	76.6	79.4	
Total		84.7	75.7	73.6	

3.4.7 Summary

The classification experiment reveals that speakers can make the best end-of-utterance classifications for bimodal, AUDIOVISUAL stimuli. It is interesting to observe that the numerically lowest scores are obtained for the AUDIO-ONLY condition, which has received most attention in the literature. The VISION-ONLY results are somewhat better, which shows that visual cues to end-of-utterance are indeed useful for participants. Besides the modality effects, some other interesting results were obtained. A small response bias was found for NON-FINAL fragments, so that NON-FINAL fragments are slightly more often classified correctly. For the NON-FINAL fragments, the LONG stimuli evoked more correct answers than the SHORT stimuli, while for the FINAL fragments the stimulus length makes almost no difference. Finally, the classification scores were found to vary per speaker, both overall and as a function of modality.

3.5 Observational analysis

The focus in this chapter has been on a perceptual comparison of the cue value of different modalities for signaling end-of-utterance. However, it would be interesting to see which auditory and visual behaviors might have served as cues in both experiments. To gain some insight into this, we annotated for both the final and the non-final stimuli, the 50% that received the best classification scores in experiment 2. In particular, we concentrated on those cues that are known from the literature (see section 3.1), and that could clearly and consistently be determined on the basis of visual or auditory inspection of our stimuli.

The following *auditory* cues were labeled:

- *Final pitch level*: whether a fragment ends in a LOW (L), MEDIUM (M) or HIGH final pitch level (H).
- *Creaky voice*: whether a stimulus contains some CREAKY fragments.

In both cases, the annotation was determined by perceptual judgments, and performed by professional intonologists. The distinction between high, mid, and low final pitch levels was determined by comparing the tonal pattern in the final syllables of the fragment to the pitch range of the preceding part. If the final stretch of speech was clearly below or above the preceding pitch range, it would be categorized as either LOW or HIGH, whereas a pitch in between those two extremes would get a MEDIUM label.

In the *visual* domain, the following features were labeled (Table **15** contains representative stills for each of the visual features):

- *Brows*: whether the eyebrows are raised (UP) or lowered (DOWN).
- *Eyes*: whether the eyes of the speaker are turned away from the camera (AWAY), or whether the speaker returns his/her gaze towards the camera (BACK); we also labeled cases where a speaker was BLINKING.
- *Mouth*: whether the mouth at the end of the fragment is CLOSED or OPEN. Note that there was always a naturally occurring pause after the end of the auditory signal (see section 3.4.1).
- *Head*: whether the speaker turns his/her head AWAY from the camera during the answer, or moves the head BACK to the camera; moreover, we also labeled cases where the speaker makes a NODDING movement during the fragment.
- *Posture*: whether the speaker changes his/her posture AWAY from the camera, or rather moves his/her body BACK towards the camera.

The cues were always labeled blind to condition, in order to avoid circularity in their annotation.

Table **15** - Representative stills illustrating the annotated visual features. Notice that various stills contain multiple features, since cues may co-occur. For example, the female speaker with her mouth open also moves her head and eyes away

Label	*Example 1*	*Example 2*
Brows raised		
Eyes diverted		
Mouth open		
Head away		
Posture away		

Table **16** gives the overall results for the factors of interest, split by the two possible modalities, i.e. AUDITORY (final pitch levels, creaky voice) and VISUAL (brows, eyes, mouth, head, posture) as a function of fragment type (NON-FINAL or FINAL).

Table **16** - Selection of a number of annotated features; the description and examples represent the marked settings for each feature

Modality	*Feature*	*Setting*	*NF*	*F*	*Total*
Auditory	Final pitch level	H	0	6	6
		M	13	2	15
		L	3	8	11
	Creaky voice		5	5	10
Visual	Brows	Up	11	8	19
		Down	3	4	7
	Eyes	Blinking	7	12	19
		Away	23	8	31
		Back	3	13	16
	Mouth	Open	6	2	8
		Closed	0	4	4
	Head	Nodding	12	21	33
		Away	10	4	14
		Back	1	4	5
	Posture	Away	7	6	13
		Back	0	2	2

In the AUDITORY domain, it can be observed that the MEDIUM-ending pitch is more typical for the NON-FINAL fragments, while both HIGH and LOW final pitch levels occur more often at the end of FINAL fragments. This result is in line with many previous studies which show that a clearly low pitch or a high pitch (such as in question intonation) may signal the end of an utterance, whereas a medium pitch serves to cue continuity (e.g. Caspers, 1998; Silverman & Pierrehumbert, 1990). At first sight, the presence of a CREAKY voice (which in our stimuli rarely happens in the first place) does not appear to be related to finality or non-finality, but a closer inspection of the stimuli revealed us that most of the CREAKY fragments co-occurred when speakers produce a HIGH or LOW pitch (in both final and non-final fragments), while in cases where speakers used a MEDIUM pitch the fragments were non-creaky (not shown in the table). Thus, creakiness may serve as an extra cue to reinforce the finality/non-finality marking of final pitch levels. With respect to the VISUAL features, Table **16** suggests that there is a clear tendency for speakers to divert their eyes and head in NON-FINAL fragments, while they return eyes, head, and also posture in the FINAL fragments. Additionally, there is a trend for the mouth to be still OPEN when a fragment has not yet been finished (even though the speaker is not speaking), whereas a mouth is more often CLOSED at the end of a FINAL fragment. Also, FINAL fragments display relatively more cases of BLINKING and NODDING[13], while the brows tend to be UP or DOWN at the end of NON-FINAL versus FINAL fragments, respectively.

There are also many individual differences between speakers. In the annotated utterances, speakers produce almost 23 cues on average, but there are clear differences. Speaker JB for instance, produces only 14 visual cues to signal finality, which is consistent with the fact that speaker JB was the most difficult to classify in experiment 2. On the other hand, speaker JB tends to use low final pitch levels more often than other speakers. This may account for the observation that, for experiment 1, participants took relatively long to respond to JB's stimuli in the VISION-ONLY modality, and were rather quick for this speaker in the AUDIO-ONLY and AUDIOVISUAL condition. Speaker SS, to give a second example, is visually the most expressive (33 visual cues) and indeed her stimuli lead to the overall quickest responses in experiment 1, and to the most correct classifications in experiment 2.

Apart from the fact that some speakers display more cues than others, some speakers also tend to display different cues than other speakers. For example, on the visual level, while most speakers return their gaze in a final position, some speakers (e.g. ED) do not return their gaze but instead nod more often in final position.

This small scale annotation reveals that many of the cues mentioned in the introduction indeed occur in the stimuli, and it seems likely that participants made their classification on the basis of these various cues. In future research, it would be interesting to find out how the different audiovisual features discussed above are distributed over the whole utterance. It has been argued (Argyle & Cook, 1976, pp.116-118; see also Kendon, 1967) that an

[13] After visual inspection of the non-final samples in which nodding takes place, we got the impression that when speakers do nod on non-final words, they have a tendency to do this in a regular pattern, in which they nod rythmically on each word, including the last word. The direction of the nods can alternate, or they nod repeatedly in the same direction except for the last word. This tendency to repeat the same behavior, such as nods, is also described by Graf et al. (2002). We did not annotate these patterns.

utterance consists of different phases, i.e. a starting phase, a middle phase, and a closing phase, which are connected to patterns in eye gaze (see also Cassell et al., 2001, for similar kinds of observations in other bodily gestures). It remains to be seen whether such patterns are also true for other visual features, and how these relate to more global auditory cues, such as declination or rhythmic patterns (but see e.g. Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). It would also be interesting to test the relative importance of the various auditory and visual cues in follow-up experiments.

3.6 Discussion and conclusion

The fact that speakers use auditory cues (intonation, pausing, rhythm etc.) which indicate that they are approaching the end of their utterance is well established (e.g. de Pijper & Sanderman, 1994; Price et al., 1991; Swerts, Bouwhuis et al., 1994; Swerts, Collier et al., 1994; Wightman et al., 1992). Various researchers have pointed out that speakers may also employ visual cues (such as posture, head movements or gaze) for this purpose (e.g. Argyle & Cook, 1976; Cassell et al., 2001; Nakano et al., 2003; Vertegaal et al., 2000). While the auditory cues have been studied from a perceptual perspective as well, comparable studies addressing the perception of visual cues (or the audiovisual combination) for end-of-utterance detection are thin on the ground. This naturally raises the question which modalities people employ to determine whether a speaker is at the end of an utterance, and what the effect is of combining information from different modalities. In order to answer this question, we first collected utterances in a semi-spontaneous way using a new experimental paradigm eliciting target list-answers of three or five words long, making sure that target words could occur at the beginning, middle or end of the list. On the basis of these utterances, two perception experiments were carried out.

As a first exploration, we performed a reaction time experiment in which participants were confronted with utterances, taken out of their original interview context to make sure that participants could not rely on lexical cues, and presented in three formats: VISION-ONLY (VO), AUDIO-ONLY (AO) or AUDIOVISUAL (AV). The task for participants was to indicate as soon as possible when the speaker reached the end of his or her current utterance. It was found that participants could do this most quickly in the bimodal, AUDIOVISUAL condition, followed (with a relative small, non-significant margin) by the AUDIO-ONLY condition, and with the slowest responses in the VISION-ONLY condition.

To find out how participants respond to stimuli in the respective conditions without any cues that participants might relate to (non)-finality, we also performed a baseline reaction time measurement using artificially created static stimuli. Even though these artificial stimuli are out of necessity not fully comparable with the real, experimental stimuli, comparing the experimental scores with those obtained in the baseline reveals some suggestive differences. It is interesting to observe that in the baseline condition, the AUDIOVISUAL stimuli led to the slowest responses. That RT's for the AV condition are relatively slower in the baseline than in the actual experiment may be explained by the thesis that when two different modalities (which contain no cues when their presentation will end) are offered at the same time, they will produce a cognitive overload because two sources of information have to be processed instead of one (Doherty-Sneddon et al., 2001). However, when two modalities are presented in a situation where the information does contain predictive cues, as in the non-baseline condition, the different modalities might serve as sources providing complementary information, and thus can help each other in resolving ambiguous slots in the stream of speech (compare Kim, Davis, & Krins, 2004; Schwartz, Berthommier, & Savariaux, 2004).

In general, the responses to the baseline stimuli were substantially faster than the responses in the non-baseline conditions. This is in line with various reaction time studies concluding that a complex stimulus leads to slower reaction times (e.g. Brebner & Welford, 1980; Luce, 1986; Teichner & Krebs, 1974). Since the baseline stimuli are essentially static, without any variations that might be informative for end-of-utterance detection, there is much less information to process than in the experimental stimuli.

It was also interesting to see that the 5 WORD stimuli lead to quicker responses than the 3 WORD ones, which is in line with the studies of Carlson et al. (2005) and Swerts & Geluykens (1994) mentioned in the introduction. Again, this result is also consistent with findings from the literature on reaction time studies. Froeberg (1907), for instance, already found that longer visual stimuli elicit faster reaction times than stimuli of a shorter duration, and Wells (1913) found the same for auditory stimuli. In general, it is known that stimulus duration has a clear impact on reaction times (e.g. Ulrich, Rinkenauer, & Miller, 1998). Moreover, in this particular set-up, the 5 WORD stimuli may also simply contain more potential finality cues than the 3 WORD stimuli, which would be an additional explanation for the fact that 5 WORD stimuli result in quicker responses than 3 WORD ones.

The results from the first experiment cannot be used to rule out the possibility that auditory information is sufficient for end-of-utterance detection, since it did not result in a significant difference between the AUDIOVISUAL and the AUDIO-ONLY condition. Therefore a second experiment was conducted, to get more insight in how participants respond to stimuli in the different modalities. In this experiment participants were offered SHORT (1 word) and

LONG (2 word) fragments which either did or did not mark the end of an utterance, and participants had to classify these as FINAL or non-FINAL. In this experiment the bimodal presentation format gave significantly better results than the unimodal ones: when participants have access to both auditory and visual cues they make more adequate classifications than in situations where they only have information from one modality at their disposal. It was interesting to observe that overall most mistakes are made in the AUDIO-ONLY condition, i.e. the situation which has received most attention in the literature so far, although the difference between the respective unimodal conditions was not statistically significant. Two possible explanations can be given for the superiority of the AUDIOVISUAL stimuli in this particular experiment.

First, a combined AUDIOVISUAL presentation format clearly offers more cues than a presentation in a single modality. But we have also seen that speakers differ in which signals they display, with some speakers showing more visual cues and other more auditory ones. Clearly, this also speaks in favor of a bimodal presentation.

In addition a slight response bias was found for NON-FINAL fragments, with NON-FINAL fragments more often classified correctly than the FINAL ones. And for the NON-FINAL fragments, it was found that the LONG stimuli were more often classified correctly than the SHORT ones, while stimulus length did not have an effect for the final fragments. This suggests that when finality cues are available, it makes no difference whether the fragment is short or long, but when finality cues are not available, participants need longer fragments to make a decision. This could be caused by the fact that finality is displayed in local cues, thus in the last part of a fragment, just before it stops. In contrast, when no local finality cues are displayed, people need to base their decision on global cues. In general, it is a well-known finding in cognitive psychology that it is easier to determine whether a cue is present than to decide that something is not there (e.g. Hearst, 1991).

It is also noteworthy that the LONG fragments are better classified than the SHORT fragments in the AUDIO-ONLY condition, which suggests that the finality cues in speech seem to be more global in nature, and hence that participants can make better judgments for longer fragments when more of these global cues are available. For the VISION-ONLY condition, length does not appear to have an influence, which suggests that the visual cues may be more local.

Notice that this would also offer an explanation for the fact that the AUDIO-ONLY condition outperforms the VISION-ONLY condition in experiment 1, but not in experiment 2. Since the stimuli in the second experiment where overall shorter fragments (consisting of 1 or 2 words) than those in the first experiment (which consisted of entire utterances of 3 or more words), the participants in the second experiment could not use the spoken global cues to full effect.

In sum: our study, using a reaction-time experiment and a classification task, has revealed that participants are sensitive both to auditory and visual signals when they need to estimate whether or not a speaker utterance has ended. While both modalities separately contain cues that enable participants to make reliable finality judgments, it turns out that a bimodal, AUDIOVISUAL condition leads to the most accurate results. The relative cue value of the two unimodal conditions depends on the experiment, where auditory cues were more important in the RT experiment, and visual cues in the classification task. In addition, its relative importance also differs between stimuli from different speakers, due to the fact that some speakers display more auditory cues, and others more visual ones.

Appendix

This appendix lists the questions in the quiz presented in this chapter. The experimental questions consists of two types of questions: generating sets of concrete words such as colors or names of objects, or generating numbers. The questions were asked in Dutch. The order of an answer to one the number questions was always fixed. For the general questions the order can vary. The answers of one of the participant (MP) are given. Wrong answers or answers with a finality cue (such as "and") were not used in the perception studies.

General questions

	Dutch	*English*
Q:	Noem vier soorten insecten	List 4 types of insects
A:	Ehm.. een mug.. een bij... een mier... een pissebed [gelach]	Uhm... a mosquito... a bee... an ant... a woodlouse [laughter]
Q:	Welke talen spreek je?	Which languages do you speak?
A:	Eh. Nederlands. Engels.. Frans.. Spaans	Uh. Dutch. English.. French.. Spanish
Q:	Wat zijn de landen in de Benelux?	What are the countries in the Benelux?
A:	Nederland Duitsland. België	The Netherlands Germany. Belgium
Q:	Noem drie landen die de Euro gebruiken	List 3 countries that use the Euro
A:	België Duitsland. Nederland	Belgium Germany. The Netherlands
Q:	Noem vier soorten groentes	List 4 types of vegetables
A:	aubergine courgette. tomaat	Eggplant zucchini. tomato
Q:	Wat zijn de kleuren van de Nederlandse vlag?	What are the colors of the Dutch flag?
A:	Rood. wit blauw	Red. white blue
Q:	Noem vier huisdieren	List 4 pets
A:	Uh poes vis. kat.. oh nee [zelfde].. hond.. en:.. cavia	Uh puss fish cat.. oh no [same].. dog.. and:.. guinea-pig
Q:	Noem drie landen in Europa waar je kunt skien	List 3 countries in Europe where you can go for skiing
A:	Italië. Frankrijk.. Oostenrijk	Italy. France.. Austria
Q:	Noem drie Europese talen	List 3 European languages
A:	Nederlands Frans. Duits	Dutch French. German

Q:	Noem drie populaire vakantielanden in Europa	List 3 countries in Europe which are popular for holidays
A:	Frankrijk. Spanje.. en:.. Griekenland	France. Spain.. and:.. Greece
Q:	Noem drie primaire kleuren	List 3 primary colors
A:	Groen. rood. blauw	Green. red. blue
Q:	Wat zijn drie kleuren die een oog kan hebben?	What are 3 possible colors of an eye?
A:	Ehm.. [drie kleuren]. groen.. blauw. en rood	Uhm.. [three colors]. green.. blue. and red
Q:	Wat zijn de kleuren van de franse vlag, in de juiste volgorde?	What are the colors of the french flag, in the right order?
A:	.Rood wit blauw	.Red white blue
Q:	Noem vier soorten fruit	List 4 types of fruit
A:	Appel banaan. kiwi.. aardbei	Apple banana. kiwi.. strawberry
Q:	Wat is je lievelingskleur?	What is your favorite color?
A:	Blauw	Blue
Q:	Noem drie wereldtalen	List 3 world languages
A:	Engels Spaans Frans.. e::n... Duits	English Spanish French.. a::nd... German
Q:	Noem drie haarkleuren	List 3 possible hair colors
A:	E:h. blond. bruin. zwart	U:h. blond. brown. black
Q:	Wat zijn de kleuren van de duitse vlag?	What are the colors of the german flag?
A:	[gelach wacht even].. rood zwart geel	[laughter wait a moment].. red black yellow
Q:	Noem drie talen die men spreekt in de Benelux	List 3 languages that are spoken in the Benelux
A:	E::hm.. Duits. Belgisch.. of.. Frans en Nederlands	U::hm.. German. Belgian.. or.. French and Dutch

Number questions

	Dutch	English
Q:	Noem alle veelvouden van 4 tussen 1 en 13, in oplopende volgorde.	List all multiples of 4 between 1 and 13, in ascending order
A:	Eh.. vier. acht.. twaalf	Uh.. four. eight.. twelve
Q:	Geef de oneven getallen tussen 10 en 0, van hoog naar laag.	List the odd numbers between 10 and 0, from high to low
A:	Eh. negen zeven vijf.. drie.. één	Uh. nine seven five.. three.. one
Q:	Noem alle veelvouden van 5 tussen 4 en 26, in oplopende volgorde	List all multiples of 5 between 4 and 26, in ascending order
A:	Vier en zesentwintig.. eh.. v:ijf. tien. vijftien.. twintig.. vijfentwintig	Four and twenty-six.. uh.. f:ive. ten. fifteen.. twenty.. twenty-five
Q:	Noem alle veelvouden van 3 tussen 8 en 16, in oplopende volgorde	List all multiples of 3 between 8 and 16, in ascending order
A:	..E:h. negen. twaalf.. vijftien	..U:h. nine. twelve.. fifteen
Q:	Geef de getallen tussen 10 en 4, van hoog naar laag	List the numbers between 10 and 4, from high to low
A:	..Tien negen acht.. zeven zes. vijf vier	..Ten nine eight.. seven six. five four
Q:	Geef de oneven getallen tussen 4 en 14, in oplopende volgorde	List the odd numbers between 4 and 14, in ascending order
A:	..E:h vijf. zeven:. negen:.. elf. dertien	..U:h five. seven:. nine:.. eleven. thirteen
Q:	Noem alle veelvouden van 12 tussen 1 en 40, in oplopende volgorde	List all multiples of 12 between 1 and 40, in ascending order
A:	..E:h. twaalf. vierentwintig.. zesendertig... mm dat was het	..U:h. twelve. twenty-four.. thirty-six... mm that was it
Q:	Geef de getallen tussen 8 en 2, van hoog naar laag	List the numbers between 8 and 2, from high to low
A:	Ehm. acht... acht zeven zes vijf vier drie twee	Uhm. eight... eight seven six five four three two
Q:	Tel af van 25 naar 4, in stappen van 5	Count down from 25 to 4, in steps of 5
A:	..Eh vijfentwintig twintig vijftien.. ja vijf	..Uh twenty-five twenty fifteen.. yes five
Q:	Noem alle veelvouden van 4 tussen 10 en 21, in oplopende volgorde	List all multiples of 4 between 10 and 21, in ascending order
A:	..Mm tien veertien. achttien.. tweeentwintig.. zesentwintig	..Mm ten fourteen. eighteen.. twenty-two.. twenty-six
Q:	Noem alle veelvouden van 6 tussen 1 en 20, in oplopende volgorde	List all multiples of 6 between 1 and 20, in ascending order
A:	..E:hm.. zes. twaalf.. achttien	..U:hm.. six. twelve.. eighteen

Q:	Noem alle even getallen tussen 7 en 13, in oplopende volgorde	List all even numbers between 7 and 13, in ascending order
A:	[mm zestien].. acht. tien. twaalf. veertien.. acht en twaalf he?	[mm sixteen].. eight. ten. twelve. fourteen.. eight and twelve isn't it?
Q:	Welk getal heeft je voorkeur, 5 of 12?	Which number do you prefer, 5 or 12?
A:	Mm.. twaalf	Mm.. twelve
Q:	Hoeveel vingers heeft een hand?	How many fingers has a hand?
A:	Vijf	Five

4 Emotional speech

in congruent and incongruent conditions

4.1 Introduction

The last chapter discussed how audiovisual speech is used to signal the end of an utterance, which may play a role in the fluency of turn-taking. In this chapter we investigate how audiovisual emotional speech is displayed.

Facial expressions are often considered to be windows to the soul, e.g. because they are thought to reveal the emotional state of a speaker[14]. From a face, we may tell whether a person is feeling happy, sad, angry, anxious, etc. (Adolphs, 2002; Carroll & Russell, 1996; Schmidt & Cohn, 2001). However, previous research has brought to light that the emotional state of a speaker can also be derived from other modalities. In the auditory domain, it has been shown that listeners can infer the emotional state from the expression of a speaker's voice (Bachorowski, 1999; Banse & Scherer, 1996; Scherer, 2003). Scherer (2003) states that acoustic emotional expressions occur at various stages (and levels) within the communication process. There is a wealth of neurobiological evidence suggesting that the recognition of emotion is a complex process which involves the cooperation of processes across various brain structures (Adolphs, 2002; Vuilleumier & Pourtois, 2007).

While we have gained much insight into how unimodal stimuli (either auditory or visual) are processed, far less is known about the extent into which these modalities interact with each other. There is some preliminary evidence that one modality may have an effect on another one, as is, for example, clear from the fact that people are able to detect from a speaker's voice whether (s)he is showing a smile (Aubergé & Cathiard, 2003). It is very likely that the brain tends to bind information received through different modalities (referred to as *intermodal* or *cross-modal binding*) (see e.g. Ghazanfar et al., 2005), because often it

---

[14] An earlier version of this chapter will be published as Barkhuysen, P., Krahmer, E., & Swerts, M. (accepted). Cross-modal and incremental perception of audiovisual emotional speech. *Language and Speech*.

receives information simultaneously through different sensory systems but from the same distal source (Pourtois et al., 2000), especially because the "sender" tends to transmit information across different modalities (see e.g. Graf et al., 2002). Generally speaking, this multimodal integration is very useful, e.g. because input from one modality can substitute another one in deteriorated circumstances. For example, lip-reading can be useful for speech comprehension in noisy environments (Sumby & Polack, 1954, in Calvert et al., 1998), or vice versa, in darkness, auditory signals can replace visual signals (Calvert et al., 1998). Neurological studies have already brought to light what the nature is of different networks activated in different brain areas during cross-modal binding, for example when involved in audiovisual speech processing (Calvert, 2001; Calvert et al., 1998; Sekiyama, Kanno, Miura, & Sugita, 2003). In the past, the binding and interaction of different modalities has been shown very spectacularly in the so-called McGurk effect, which shows that the auditory perception of a sound can be altered by the display of incongruent visual information (McGurk & MacDonald, 1976). The McGurk paradigm has been a source of inspiration for studies on the perception of audiovisual speech and/or emotions, which use stimuli with congruent and incongruent auditory and visual cues (Aubergé & Cathiard, 2003; de Gelder & Vroomen, 2000; Hietanen, Manninen, Sams, & Rusakka, 2001). However, while much research has been done about cross-modal integration during *audiovisual speech* processing, much more needs to be done about cross-modal integration during the processing of *emotions* (e.g. de Gelder, Böcker, Tuomainen, Hensen, & Vroomen, 1999; Pourtois et al., 2000), when combined with audiovisual speech. It has been shown that the ability to integrate information from emotional faces with emotional prosody is already present in 7-month-old infants (Grossmann, Striano, & Friederici, 2006). Unfortunately, many of the studies investigating the recognition of emotional expressions have been based on analyses of static images, such as photographs or drawings (see e.g. Ekman et al., 1972, pp.49-51), rather than dynamic images. As a result, little is known about the perception of emotions through "fleeting changes in the countenance of a face" (Russell et al., 2003, p.330). Often, a realistically varying speech signal is combined with a static face, resulting in knowledge about online auditory speech but not about online visual speech. Consequently, we do not yet fully understand whether auditory and visual cues of emotional speech differ in perceptual strength, and how people deal with input coming from two modalities when they have to make judgments about a speaker's emotional state (in contrast to judging an emotional state *without* speech). This knowledge could be very useful for the development of computerized speech systems, for instance (Cohn & Katz, 1998). Therefore, the first question we want to explore in this chapter is whether the processing of emotional speech is integrated across modalities, i.e. whether the perception of a combination of two modalities is more successful than the perception of a single modality alone.

A second question that we want to explore in this chapter is to what extent the recognition of emotion varies as a function of the time that people are exposed to the facial expressions of a speaker. There are reasons to believe that this temporal recognition process may vary for different kinds of emotions, such as positive versus negative emotions. That is, it has been argued that positive and negative emotions are not recognized equally fast, although there is some controversy about the direction of this effect. Fox, Lester, Russo, Bowles, Richter & Dutton (2000) claim that angry facial expressions are detected more rapidly than happy expressions, whereas Leppänen and Hietanen (2004) report that positive facial expressions are recognized faster than negative ones[15]. Potentially, the valency effect on recognition speed, in whichever direction, may partly be due to timing-related differences in facial expressions. In addition, there is work on the time-course of intermodal binding of emotions, where it appears that integration of emotional information from the face and from the voice occurs at an early stage of processing (before both modalities have been fully processed), and uses low-level perceptual features (de Gelder et al., 1999). According to Pourtois et al. (2000), intermodal binding of emotions occurs around 110 *ms* post-stimulus, which is earlier than the processing of intermodal speech, which lies around 200 *ms* post-stimulus (Pourtois et al., 2000; see also Sekiyama et al., 2003). However, as mentioned above, this study worked with the presentation of static rather than dynamic faces. There is neurological evidence that moving faces are processed by a fundamental different path than static faces (Humphreys et al., 1993).

As mentioned above, many emotion studies rely on "acted" data. The work of Ekman (e.g. 1987, 1993), for instance, is based on posed photographs of actors, and also in speech research actors are frequently used. Additionally, many studies, in line with the McGurk paradigm, make use of stimuli that consist of incongruent cues to various emotions (e.g. conflicting visual and auditory cues). An important question is whether such stimuli are *ecologically valid*, in that acted or incongruent emotions may be more "controlled" than the spontaneous display of emotions in natural interactions. Neurological studies have shown that voluntary expressions are fundamentally different in nature from spontaneous

[15] Note that closer inspection of the stimuli used in these studies reveals that the angry stimuli in the last two experiments reported in Fox et al. (2000) are similar to the sad stimuli in the experiments of Leppänen and Hietanen (2004), basically using very similar stylized emoticons to reflect these emotions.

expressions (Gazzaniga & Smylie, 1990; Rinn, 1984; Rinn, 1991). From a corpus study, Valstar, Pantic, Ambadar & Cohn (2006) conclude that these two can be distinguished on the basis of the speed, duration and sequence of brow movements. Similarly, there is some work into timing-related differences between spontaneous and posed smiles (also known as Duchenne and non-Duchenne smiles (see Ekman, 2004, pp.204-209, for a description; Ekman, Davidson, & Friesen, 1990)). Cohn and Schmidt (2004) report that spontaneous smiles, as opposed to posed smiles, have a smaller amplitude, have an onset that is more related to the duration (i.e. longer smiles are slower in onset), can have multiple rises of the mouth corners, and are accompanied by other facial actions, either simultaneously or immediately following.

In sum, the aim of this chapter is to look into more detail at the perception of audiovisual expressions of positive and negative emotions (both congruent and incongruent) in spoken language, and to explore the recognition speed of these dynamic expressions of positive and negative emotions (both congruent and incongruent). It describes two perception experiments and an observational study for which we used Dutch data collected via a variant of the Velten technique. This is an experimental method to elicit emotional states in participants, by letting speakers produce sentences increasing in emotional strength (Velten, 1968). The next section first describes previous work by Wilting, Krahmer & Swerts (2006), whose data were used in the current chapter. We present a brief summary of their method and the results of an experiment in which they first elicit congruent and incongruent emotional data from speakers using an adaptation of the Velten technique, and then selected film clips (without sound) which they showed to observers who had to judge the emotional state of the recorded speakers. The later sections describe how the current study uses the data collected by Wilting et al.'s research by testing these experimental stimuli in both bimodal and unimodal conditions. For reasons described below, the participants in the current study were native speakers of Czech, who were not able to understand the lexical content of the presented utterances. In the second experiment we test the original experimental stimuli (but presented without sound) on Dutch participants using a *gating paradigm* (Grosjean, 1996). Our final study consists of observational analyses of various facial expressions in the upper and lower areas of a speaker's face to see whether certain features correlate with reported or perceived emotions from speakers.

4.2 Audiovisual recordings

Wilting et al. (2006) used an adapted Dutch version of the original Velten (1968) induction procedure, using 120 sentences evenly distributed over three conditions (POSITIVE, NEUTRAL and NEGATIVE)[16]. Besides the three conditions described by Velten for the induction of congruent emotions (POSITIVE, NEUTRAL, NEGATIVE), two "acting" conditions were added. In one of these, participants were shown negative sentences and were asked to utter these as if they were in a positive emotion (INCONGRUENT POSITIVE); in the other, positive sentences were shown and participants were instructed to utter these in a negative way (INCONGRUENT NEGATIVE). The sentences showed a progression, from neutral ("Today is neither better nor worse than any other day") to increasingly more emotional sentences ("God I feel great!" and "I want to go to sleep and never wake up again" for the positive and negative sets, respectively), to allow for a gradual build-up of the intended emotional state.

Participants were told that the goal of the experiment was to study the effect of mood on memory recall (earlier work has revealed that mood induction procedures become more effective when the induction serves a clear purpose, e.g. Westermann, Spies, Stahl, & Hesse, 1996). The instructions, a slightly abridged version of the original instructions from Velten, were displayed on the computer screen, and participants were instructed to first silently read the texts, after which they had to read them aloud. For the congruent conditions, the participants were instructed to try to "feel" and "display" the emotion which the sentence was representing, while for the incongruent conditions, the participants were instructed to try to "feel" and "display" the opposite emotion[17].

During the data collection, the sentences were displayed on a computer screen for 20 seconds, and participants were instructed to read each sentence first silently and then out loud. Recordings were made from the face and upper body of the speakers with a digital

---

[16] We chose to classify the emotions under investigation according to their valence, i.e. positive and negative, instead of using a subjective term as 'happy' or 'depressed', because we were only interested in the valence of an emotion and not in specific properties of an individual emotion.

[17] Note that although the terminology in our instruction reflected only the valence of the emotion, the list designed by Velten should invoke the emotions "elation" and "depression" (Velten, 1968). However, these two emotions differ primarily along one dimension, i.e. positive to negative, according to the dimensional view upon emotions (e.g. Bachorowski, 1999). By instructing the participants to feel and display the opposite emotion as the one reflected in the sentences, we tried to direct the way they would "act" by the content of the list rather than by terminology.

camera, and a microphone connected to the camera. Fifty Dutch speakers (10 per condition) were recorded in the data collection, 31 female and 19 male, none of them being a (professional) actor. The advantage of using different speakers across conditions is that, in the perception tests, observers could not base their judgments upon the familiarity of the faces, therefore preventing learning effects. Some representative stills are shown in Figure **12**.

Positive **Negative**

Positive incongruent **Negative incongruent**

Figure **12** - Representative stills of congruent (top) and incongruent (bottom) emotional expressions, with on the left hand side the positive and on the right hand side the negative versions

Immediately following this phase, participants had to fill in a short mood questionnaire ("At this moment, I feel . . . ") derived from Mackie & Worth (1989) and Krahmer, van Dorst & Ummelen (2004), consisting of six 7-point bipolar semantic differential scales, using the following adjective pairs (English translations of Dutch originals: happy/sad, pleasant/unpleasant, satisfied/unsatisfied, content/discontent, cheerful/sullen and in high spirits/low-spirited). The order of the adjectives was randomized; for ease of processing negative adjectives were mapped to 1 and positive ones to 7.

Wilting et al. (2006) reported 2 main findings. First, from the survey presented to participants after the elicitation phase, it turned out that the Velten technique was very effective in that the positive and negative emotions could indeed be induced through this method, but only for speakers in the congruent conditions; the speakers in the incongruent conditions did not feel different from the speakers in the neutral condition. Second, observers turned out to be able to reliably distinguish between positive and negative emotions on the basis of visual cues; interestingly, the incongruent versions led to more extreme scores than the congruent ones, which suggests that the incongruent emotions were displayed more strongly than the congruent ones.

In this chapter, we are interested in the question in what sense the *positive* emotions differ from their negative counterparts. We investigate the hypothesis that one difference is *durational*, especially in the onset, assuming that positive emotions appear quicker on the face than negative ones, though this may be different for congruent versus incongruent emotions. Also, we are interested in the question whether the perception of positive versus negative emotions differs across *modalities*, and whether the perception of *congruent* versus *incongruent* emotions differs across modalities, and/or whether there is an interaction between these two. In the next study we test these data in both bimodal and unimodal conditions, on Czech participants.

4.3 Experiment 1: Classification

4.3.1 Stimuli

From each of the speakers in the recordings, the last sentence was selected. These sentences captured the speakers at the maximum height of the induced emotion. We chose to use maximum height stimuli, because Horstmann (2002) reported that prototypical emotions resemble the most intense expression of an emotion. The previous study by Wilting et al. (2006) was conducted with VISION-ONLY stimuli presented to Dutch participants. It would not have been possible to present the AUDIO-ONLY or AUDIOVISUAL variants to Dutch participants, as the lexical information would be a give away clue for the speaker's emotional state. Still, we are interested in the perception of the AUDIO-ONLY and AUDIOVISUAL stimuli.

Therefore the Dutch sentences were presented to Czech participants in the perception test, as they did not understand Dutch.

4.3.2 Design

The experiment uses a repeated measurements design with modality as between-subjects factor (with levels: AUDIOVISUAL: AV, VISION-ONLY: VO and AUDIO-ONLY: AO), condition as within-subjects factor (with levels: INCONGRUENT NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE and INCONGRUENT POSITIVE), and perceived emotional state as the dependent variable.

4.3.3 Procedure

Participants were told that they would see or hear 50 speakers in different emotional states, and that their task was to rate the perceived state on a 7 point valency scale ranging from 1 (= very negative) to 7 (= very positive). Participants were not informed about the fact that some of the speakers were displaying an incongruent emotion. Within each modality, there were two subgroups of participants, who were presented with the same stimuli but in a different random order to compensate for potential learning effects. Stimuli were preceded by a number displayed on the screen indicating which stimulus would come up next, and followed by a 3 second interval during which participants could fill in their score on an answer form. Stimuli were shown only once. The experiment was preceded by a short training session consisting of 5 stimuli of different speakers uttering a non-experimental sentence to make participants acquainted with the stimuli and the task. If all was clear, the actual experiment started, after which there was no further interaction between the participants and the experimenter. The perception tests in the three conditions were conducted as a group experiment with the material presented on a large screen in front of the class room. The entire experiment lasted approximately 10 minutes.

4.3.4 Participants

Fifty-four people (18 per condition) participated in the experiment, 9 female and 45 male, with an average age of 23 (range 21-30). All were students and PhD-students from the Czech Technical University (Faculty of Electrical Engineering) and the Charles University (Faculty of Philosophy and Arts) in Prague, Czech Republic. The choice for Czech participants was arbitrary; the only real constraint was that the participants could not understand Dutch.

4.3.5 Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse-Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. *Post hoc* analyses were performed with the Bonferroni method.

4.3.6 Results

Figure **13** and Table **17** summarize the results. A repeated measures analysis of variance (ANOVA), with modality as between-subjects factor, condition as within-subjects factor, and perceived emotional state as the dependent variable, shows that *condition* has a significant effect on perceived emotional state ($F(4, 204) = 145.042$, $p < .001$, $\eta^2_p = .740$). Repeated contrasts revealed that all conditions (level 1: INCONGRUENT NEGATIVE, level 2: NEGATIVE, level 3: NEUTRAL, level 4: POSITIVE and level 5: INCONGRUENT POSITIVE) lead to a significantly different perceived emotion ($F_{12}(1, 51) = 89.558$, $p < .001$, $\eta^2_p = .637$; $F_{23}(1, 51) = 50.167$, $p < .001$, $\eta^2_p = .496$; $F_{34}(1, 51) = 43.855$, $p < .001$, $\eta^2_p = .462$; $F_{45}(1, 51) = 20.052$, $p < .001$, $\eta^2_p = .282$). It is interesting to observe that the *incongruent* emotions are perceived as more intense than the congruent ones. Speakers in the INCONGRUENT POSITIVE condition are overall perceived as the most positive ($M = 4.70$, $SD = 0.53$), and speakers in the INCONGRUENT NEGATIVE condition are perceived as the most negative ($M = 2.72$, $SD = 0.63$). Note that the perceptual difference between incongruent and congruent emotional speech is larger for the *negative* emotions. In general, it seems that the incongruent emotions are classified "better", or interpreted as more intense than the congruent emotion.

Table **17** - Perceived emotional state on a 7-point scale (1 = very negative, 7 = very positive) as a function of condition (standard deviations between brackets) as well as condition split by modality

Condition	*AV*	*VO*	*AO*	*Total*
Inc. pos.	4.69 (.35)	4.84 (.35)	4.57 (.78)	4.70 (.53)
Positive	4.66 (.46)	4.78 (.46)	3.86 (.95)	4.43 (.77)
Neutral	3.54 (.31)	3.57 (.46)	4.42 (.49)	3.84 (.59)
Negative	3.08 (.49)	3.28 (.47)	3.54 (.77)	3.30 (.61)
Inc. neg.	2.38 (.36)	2.99 (.64)	2.79 (.72)	2.72 (.63)
Total	3.67 (.98)	3.89 (.91)	3.84 (.98)	3.80 (.96)

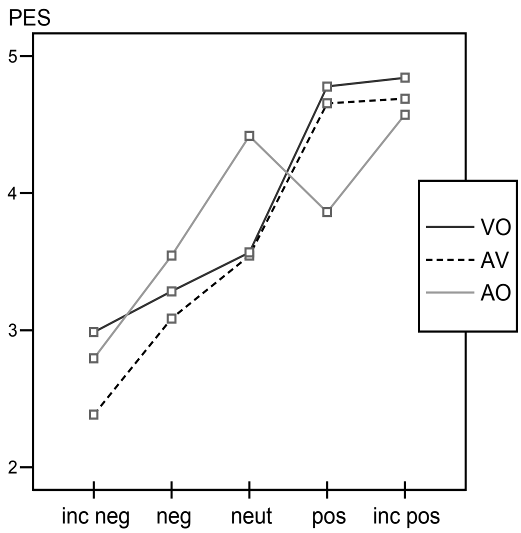

Figure **13** - The mean perceived emotional state (1 = very negative, 7 = very positive) per condition and modality

Modality does not have a significant main effect on perceived emotional state ($F(2, 51) = 1.881$, $p = .163$, $\eta^2_p = .069$), but interestingly there was an interaction between *condition* and *modality* ($F(8, 204) = 10.981$, $p < .001$, $\eta^2_p = .301$). In all three modalities the *incongruent* emotions are perceived as more intense than the congruent ones; speakers in the INCONGRUENT POSITIVE condition are perceived as the most positive, and speakers in the INCONGRUENT NEGATIVE condition are perceived as the most negative. However, repeated contrasts showed that all levels of condition and modality interact significantly with each other ($F_{12}(1, 51) = 5.438$, $p < .01$, $\eta^2_p = .176$; $F_{23}(1, 51) = 5.254$, $p < .01$, $\eta^2_p = .171$; $F_{34}(1, 51) = 41.526$, $p < .001$, $\eta^2_p = .620$; $F_{45}(1, 51) = 13.475$, $p < .001$, $\eta^2_p = .346$). For both the AV and the VO modality the difference between POSITIVE and INCONGRUENT POSITIVE is very small ($D_{AV} = 0.03$, and $D_{VO} = 0.06$), while this difference is much larger in the AO modality ($D_{AO} = 0.71$): for this modality, the POSITIVE condition even scored lower on the valency scale than NEUTRAL. On the other side of the spectrum, the difference between the NEGATIVE and the INCONGRUENT NEGATIVE condition is substantial for the AO and the AV modality ($D_{AO} = 0.75$, and $D_{AV} = 0.70$), but here the VO modality stands out in the sense that the difference is relatively small ($D_{VO} = 0.29$). In other words, the classification pattern for the AV modality resembles the VO modality for the positive moods, while for the negative moods the pattern of the AV modality is similar to the AO modality. Note also that the difference between the two incongruent emotions is larger in the AV modality ($D_{AV} = 2.31$), somewhat smaller in the

VO modality (D_{VO} = 1.85) and the smallest in the AO modality (D_{AO} = 1.78). Another interesting point is the difference between the facial expressions and vocal expressions in the POSITIVE condition ($D_{VO\text{-}AO}$ = 0.92). This difference is very large in comparison to the other conditions, apart from the NEUTRAL condition, where, in contrast to the POSITIVE condition, the AO modality scores higher than the VO modality ($D_{VO\text{-}AO}$ = -0.85).

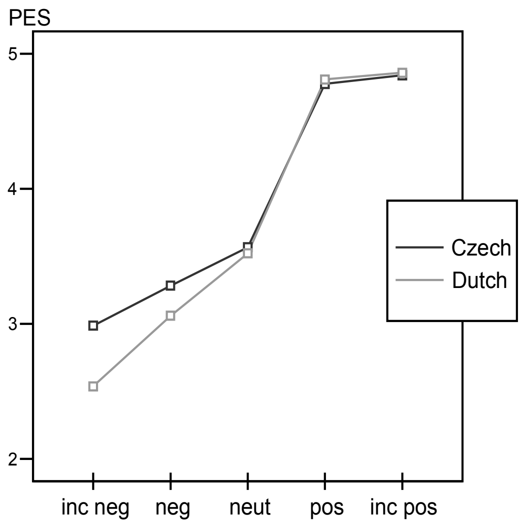

Figure **14** - The mean perceived emotional state (1 = very negative, 7 = very positive) per condition and nationality

Further, we compared the classification of the Czech participants for the fragments presented in the VO modality with the results of the earlier Dutch perception test (Wilting et al., 2006), by a repeated measures analysis of variance (ANOVA), with nationality as between-subjects factor, condition as within-subjects factor, and perceived emotional state as the dependent variable. It turns out that the main effect of *nationality* was not significant ($F(1, 56) = 1.905$, $p = .173$, $\eta^2_p = .033$). There was a significant interaction between *nationality* and *condition* ($F(4, 224) = 5.088$, $p < .01$, $\eta^2_p = .083$); however, repeated contrasts showed that this difference was only caused by the difference between the NEGATIVE and the INCONGRUENT NEGATIVE stimuli (F_{12} (1, 56) = 4.505, $p = .038$, $\eta^2_p = .074$).

4.3.7 Summary

We have reported on a perception experiment in which Czech participants rated their perceived emotional state of Dutch speakers. These speakers could either display a positive or a negative emotion, which was either congruent or incongruent. The Czech participants were confronted with these utterances in a bimodal (AUDIOVISUAL) or a unimodal (AUDIO-ONLY or VISION-ONLY) condition.

There was no overall effect of modality. Further, it was found that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, where the difference between incongruent and congruent emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between incongruent and congruent emotions were perceived in the AUDIO-ONLY condition, which suggests that displaying an incongruent emotion has a particularly strong effect on the spoken realization of emotions. This difference between the congruent and the incongruent conditions is in particular larger for the *positive* emotions. In addition, comparing the different modalities suggests that positive emotions are more clear in the VISION-ONLY modality (since the highest scores were obtained in the AV and VO modalities), while the classification of negative emotions in the AV modality follows the pattern of the AO modality. Another interesting point is the difference between facial and vocal expression within the separate conditions. It seems that the Velten procedure did not elicit recognizable vocal expressions in the POSITIVE condition, whereas it elicited recognizable facial expressions. On the other hand, the speakers in the INCONGRUENT POSITIVE condition were able to display recognizable facial *and* vocal expressions. We also compared the classification of the Czech participants for the VO fragments with the results of the Dutch perception test with the same stimuli (Wilting et al., 2006), which lead to essentially the same results.

Although we have shown that participants can correctly classify dynamical expressions of (congruent and incongruent) emotions, we did not investigate the speed with which these expressions were classified. This is interesting in the light of the above discussed timing differences between spontaneous and voluntary expressions. We also do not know whether there are timing differences between positive and negative emotions. The second experiment will investigate whether positive and negative emotions (both congruent and incongruent) differ with respect to the speed with which they are recognized as such.

4.4 Experiment 2: Gates

4.4.1 Stimuli

The second perception test is based on the *gating paradigm*, which is a well-known design in spoken word recognition research (Grosjean, 1996). In this paradigm, a spoken language stimulus is presented in segments which increase in length and participants are asked to propose the word being presented and to give a confidence rating after each segment. The dependent variables are the *isolation point* of the word (i.e. the *gate*[18]), the *confidence ratings* at various points in time and the *word candidates* proposed after each segment.

The current perception test resembles this gating design, but only in that we present parts of the original sentences used in Wilting et al. (2006), increasing in length. To enable comparisons across experiments, the fragments were cut from the start of the original fragment as it was used in experiment 1. The first segment is very short, only consisting of 4 frames (160 *ms*). The size of the later segments increases in steps of 160 *ms* until the last, sixth segment which is 960 *ms* long. Each segment S+1 thus includes the preceding segment S, and extends it by 4 extra frames (or 160 extra *ms*). We only used 6 segments, because a pilot study indicated that adding longer segments did not lead to a substantial increase in recognition accuracy.

The current set-up differs from the "standard" gating approach, in that we do not ask participants to give confidence ratings. Rather, after each gate, participants have to indicate whether they believe that the speaker is in a positive or in negative mood, or whether they cannot make this distinction on the basis of the current gate.

4.4.2 Design

The experiment uses a repeated measurements design with condition (with levels: INCONGRUENT NEGATIVE, NEGATIVE, POSITIVE and INCONGRUENT POSITIVE) and gate (with levels: ONE (i.e. 160 *ms*), TWO (i.e. 320 *ms*), THREE (i.e. 480 *ms*), FOUR (i.e. 640 *ms*), FIVE (i.e. 800 *ms*), to SIX (i.e. 960 *ms*)) as within-subjects factors, and confidence (with levels: NON-ANSWERS "don't know" versus ANSWERS "positive or negative") and perceived emotional state (with levels: POSITIVE and NEGATIVE) as the dependent factors.

---

[18] In our perception test, the isolation point is rather the gate at which a fragment is correctly recognized and where responses for following gates are no longer changed.

4.4.3 Procedure

Participants were tested individually. They were invited into a quiet room, and asked to take place in front of the computer. Participants were told that they would see 40 speakers in different emotional states, and that for each speaker they would see 6 short, overlapping fragments (the gates). The task of the participants was to determine, for each gate, whether the speaker was in a positive or in a negative mood. They were given 3 answering possibilities: "negative", "don't know", and "positive". Three keys on the keyboard were labeled with these answer possibilities, and *only after* viewing a film clip, participants could press one of these buttons, after which the next stimulus appeared. Therefore, they could take as much time as they needed for judging the film clip, while they were viewing a blank screen. However, the instruction encouraged the participants to respond quickly. If they were not sure yet about the emotion of the clip, they could use the "don't know" button, which was designed for this purpose. Participants were not informed about the fact that some of the speakers were displaying an incongruent emotion.

The gates were presented in a *successive format*: that is, participants viewed all the segments of a sentence, starting with the shortest and finishing with the longest. The gates were presented *forwards*, i.e. the first was cut from the beginning of the sentence and then increasingly longer stretches were added, thus later segments were approaching the end ("left-to-right"). Stimulus groups (containing six gates) were preceded by a number displayed on the screen indicating which stimulus group would come up next, and followed by the first segment only after which the participants could press the appropriate button to indicate their answers. Stimuli were shown only once. Stimulus groups were presented in one of four random orders, to compensate for potential learning effects. The fragments were only presented visually, without the corresponding sound; therefore the lexical or grammatical content could not influence the participants' decision. Also, no feedback was given to participants about the correctness of their scores.

The experiment was preceded by a short training session consisting of 1 stimulus group containing 6 gates, uttered by a single speaker uttering a non-experimental, neutral sentence to make participants acquainted with the stimuli and the task. If all was clear, the actual experiment started, after which there was no further interaction between the participants and the experimenter. The entire experiment lasted approximately 25 minutes.

4.4.4 Participants

Forty people (10 per presentation order) participated in the experiment, 33 female and 7 male, with an average age of 19 (range 18-27). All were students from Tilburg University in The Netherlands, none had participated as a speaker in the study by Wilting et al. (2006) or in experiment 1, and all were unaware of the experimental question.

4.4.5 Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse-Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. *Post hoc* analyses were performed with the Bonferroni method.

4.4.6 Results

We report on the results in two steps, first we look at the percentages of ANSWERS and NON-ANSWERS as a function of gate, and next we look at the number of POSITIVE and NEGATIVE answers as a function of gate.

First of all, we present the general distribution of responses across the conditions in Table **18**[19].

[19] There seems to be a response bias towards negative responses, i.e. the number of "positive responses" for the positive and the incongruent positive conditions is higher than the number of "don't know" responses. Therefore, within these conditions, the mean perceived emotional state "drops" in the later gates. This could be caused by the successive forward presentation format. According to Grosjean (Craig & Kim, 1990, and Walley et al., 1995, in 1996), in this design potential artefacts may occur: "The successive presentation format may induce response perseveration and negative feedback. This in turn may yield a slightly conservative picture of recognition". However, the tendency for less extreme or more negative responses in the positive condition is in line with the results of Wilting et al. (2006) and with the results in the first perception experiment. Therefore, we do not consider this to be a problem.

Table **18** - Perceived emotional state as a function of condition (standard errors between brackets) as well as condition split by gate

Gate	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
Inc. pos.	0.81 (.03)	0.77 (.03)	0.74 (.02)	0.74 (.02)	0.75 (.02)	0.75 (.02)	0.76 (.02)
Positive	0.76 (.03)	0.67 (.03)	0.64 (.02)	0.64 (.02)	0.64 (.02)	0.64 (.02)	0.67 (.02)
Negative	0.26 (.04)	0.23 (.03)	0.25 (.02)	0.26 (.03)	0.23 (.02)	0.22 (.03)	0.24 (.02)
Inc. neg.	0.20 (.03)	0.14 (.02)	0.14 (.03)	0.15 (.03)	0.14 (.03)	0.13 (.02)	0.15 (.02)
Total	0.51 (.02)	0.46 (.01)	0.44 (.01)	0.45 (.01)	0.44 (.01)	0.44 (.01)	

Non-answers versus answers

For this analysis, we recoded the responses such that NON-ANSWERS ("don't know") were mapped to a value of 0 (=no decision made), and ANSWERS ("negative" or "positive") were mapped to 1. There were 1112 NON-ANSWERS, which is 11.6% of all responses. There were a total of 191 missing values, which is 2% of all responses; these were replaced with the mean value over the 10 speakers per gate. Figure **15** shows the proportion of ANSWERS as a function of gate. We assumed that the proportion of ANSWERS is a reflection of the level of confidence that the participants have in their ability to make a correct judgment at that particular gate. What this figure shows is that we find the most NON-ANSWERS for the first gate, and that the *congruent* emotions get more NON-ANSWERS than their incongruent counterparts. In all conditions, the percentage of ANSWERS increases over the next gates, and seems to reach a plateau after the fourth gate (640 *ms*). Also, the speed of recognition (i.e. how much visual information, defined as the number of gates, is needed) differs for *positive* versus negative emotions. Taking an 80% threshold[20], it can be seen that the recognition of *positive* emotions reaches this level already at gate 2 (=320 *ms*) (congruent: M = 0.83, SE = 0.028; incongruent: M = 0.87, SE = 0.025), while the *negative* emotions reach this level only at gate 3 (=480 *ms*) (congruent: M = 0.87, SE = 0,031; incongruent: M = 0.87, SE = 0.026).

[20] Grosjean (Tyler & Wessels, 1983, in 1996) reports about a study that used this threshold as a recognition point, although there is no consensus about which threshold reflects the 'real' recognition point.

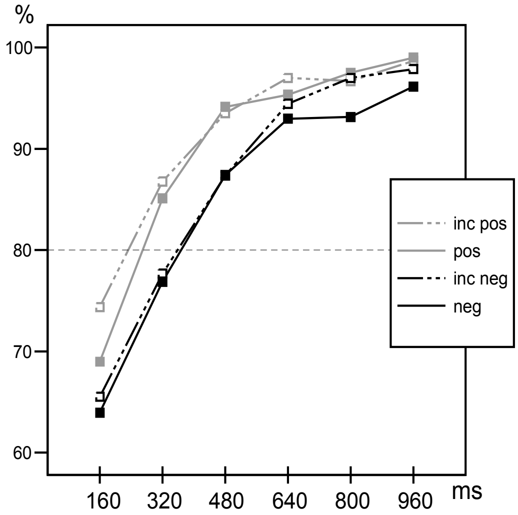

Figure **15** - The mean proportion of answers (vs. non-answers) as a function of gate (in *ms*) for different emotions

A repeated measures analysis of variance with condition and gate as within-subjects factors and proportion of answers (i.e. the confidence) as the dependent variable shows that condition has a significant effect on the proportion of answers ($F(3, 117) = 8.051$, $p < .001$, $\eta^2_p = .171$). *Post hoc* analyses reveal that the *positive* conditions differ from the negative ones ($p < .05$) but the *congruent* conditions do not differ significantly from the incongruent ones. The relative proportion of answers also differs across the *gates* ($F(5, 195) = 47.138$, $p < .001$, $\eta^2_p = .547$). *Post hoc* analyses reveal that all gates differ significantly from each other ($p < .01$) except gate 4 and 5 ($p = 1$). Finally there is an interaction between *condition* and *gate* ($F(15, 585) = 2.914$, $p < .01$, $\eta^2_p = .070$).

We also performed univariate analyses within a condition, with gate as within-subjects factor and proportion of answers as the dependent variable, in order to see how the relative proportion of answers differs across the gates between *positive* and negative emotions, both *congruent* and incongruent. Within the INCONGRUENT NEGATIVE condition ($F(5, 195) = 33.529$, $p < .001$, $\eta^2_p = .462$), *post hoc* analyses show that gates 1 to 4 differ significantly from each other ($p < .05$). Within the NEGATIVE condition ($F(5, 195) = 34.622$, $p < .001$, $\eta^2_p = .470$), gates 1 to 3 differ significantly from each other ($p < .001$). Within the POSITIVE condition ($F(5, 195) = 40.511$, $p < .001$, $\eta^2_p = .510$), gates 1 to 3 differ significantly from each other ($p < .01$), as well as gates 4 and 6 ($p < .05$). Within the INCONGRUENT POSITIVE condition ($F(5, 195)$

= 30.774, $p < .001$, $\eta^2_p = .441$), gates 1 to 3 differ significantly from each other ($p < .01$), as well as gates 3 and 6 ($p < .05$).

Finally, we performed univariate analyses within gate 1, with condition as within-subjects factor and proportion of answers as the dependent variable, in order to see whether the differences between conditions are present from the beginning. For gate 1 ($F(3,117) = 5.949$, $p < .01$, $\eta^2_p = .132$), *post hoc* analyses revealed that all conditions differ significantly from each other ($p < .05$) except the POSITIVE condition, which does not differ from any condition.

Perceived emotional state

For this analysis, we recoded the original responses such that the "negative" responses obtained a value of 0, and the "positive" responses obtained a value of 1. The "don't know" responses were treated the same as the missing values. All these NON-ANSWERS were subsequently replaced by the mean of the 10 presented speakers per gate. We used this strategy because the "don't know" responses were already processed in the first step of the statistical analyses. In this *successive* step we want to know whether the distribution of *positive* versus negative ANSWERS differs across the conditions for all those cases where the participants were certain about their classification and therefore *did* choose an answer. So, while the first step reflects the level of uncertainty across all responses, this step reflects the 'correctness'[21] of the ANSWERS for all the 'certain' responses. For this analysis, there was a total of 1303 NON-ANSWERS, which is 13,6% of all responses. Data are shown in Figure **16**.

[21] Therefore, this level is comparable with the variable word candidates in the standard gating paradigm.

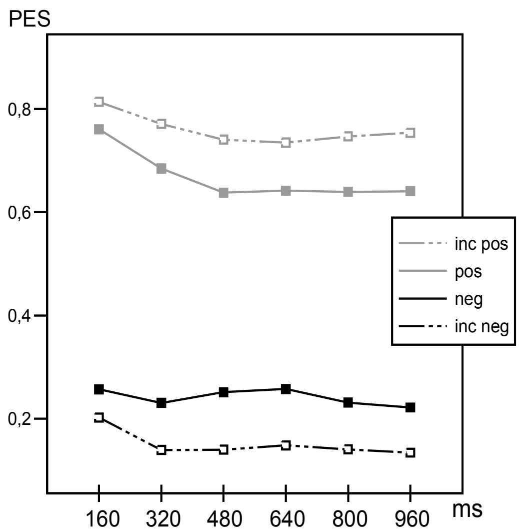

Figure **16** - The mean perceived emotional state (0 = negative, 1 = positive) as a function of gate (in *ms*) for different emotions

A repeated measures analysis of variance, with condition and gate as within-subjects factors and perceived emotional state as the dependent variable, shows that *condition* has a significant effect on the perceived emotional state ($F(3, 117) = 219.238$, $p < .001$, $\eta^2_p = .849$). *Post hoc* analyses reveal that all conditions differ significantly from each other ($p < .001$). It is interesting to observe that the *incongruent* emotions received more extreme mean classification scores than the congruent ones. Speakers in the INCONGRUENT POSITIVE condition are overall classified as the most POSITIVE ($M = 0.76$, $SE = 0.018$), and speakers in the INCONGRUENT NEGATIVE condition are classified as the most NEGATIVE ($M = 0.15$, $SE = 0.021$). The perceived emotional state also differs across *gates* ($F(5, 195) = 9.689$, $p < .001$, $\eta^2_p = .199$). *Post hoc* analyses show that only gate 1 differs significantly from all other gates ($p < .05$). Finally, there is *no* interaction between *condition* and *gate* ($F(15, 585) = 2.036$, $p = .06$, $\eta^2_p = .050$).

As with the previous tests on relative proportion of answers, we also performed univariate analyses within a condition, with gate as within-subjects factor and perceived emotional state as the dependent variable. Within the INCONGRUENT NEGATIVE condition ($F(5,195) = 3.298$, $p < .05$, $\eta^2_p = .078$), *post hoc* analyses revealed no significant differences. Within the NEGATIVE condition, only gates 4 and 6 differ significantly from each other ($p < .05$), however the overall effect of *gate* is *not* significant ($F(5, 195) = 0.867$, $p = .442$, $\eta^2_p = .022$). Within the POSITIVE condition ($F(5, 195) = 9.586$, $p < .001$, $\eta^2_p = .197$), only gate 1 differs significantly from all other gates ($p < .05$), except for gate 2, which does *not* differ significantly from any other gate[22]. Within the INCONGRUENT POSITIVE condition ($F(5, 195) = 4.736$, $p < .01$, $\eta^2_p = .108$), only gates 1 and gate 4 differ significantly from each other ($p < .05$). Therefore, it seems that in general, after gate 1, there are *no* substantial differences anymore in the classification patterns.

Because the *confidence levels* do not change substantially either in gates 4 to 6, it is interesting to look at the classification patterns within the first 3 gates. To test this, we performed a repeated measures analysis of variance, with condition and gate as within-subjects factors and perceived emotional state as the dependent variable, within the first 3 gates. Here, the effect of *condition* is again significant ($F(3, 117) = 212.042$, $p < .001$, $\eta^2_p = .845$), as well as the effect of *gate* ($F(2, 78) = 10.551$, $p < .001$, $\eta^2_p = .213$). *Post hoc* analyses showed that only gate 1 differs significantly from gate 2 and 3 ($p < .01$). So, it seems that there is a *transition point* at gate 2, which can be compared with the *isolation point* in the standard gating paradigm. There was again *no* interaction between *condition* and *gate* ($F(6, 234) = 2.261$, $p = .06$, $\eta^2_p = .055$).

Finally, because we were interested in the effect of condition within gate 1, we performed a univariate analysis of variance with condition as within-subjects factor and perceived emotional state as the dependent variable, in order to explore how participants recognize emotions within the shortest time interval. Within the first gate, the effect of *condition* is significant ($F(3, 117) = 127.729$, $p < .001$, $\eta^2_p = .766$). *Post hoc* analyses show that the *positive* conditions (i.e. the POSITIVE and the INCONGRUENT POSITIVE) differ from both negative ones ($p < .05$) but the *congruent* conditions (i.e. the POSITIVE and the NEGATIVE) do not differ from the incongruent conditions. The *positive* conditions are correctly classified as more positive (congruent: $M = 0.76$, $SE = 0.027$; incongruent: $M = 0.81$, $SE = 0.027$) and the *negative* conditions are correctly classified as more negative (congruent: $M = 0.26$, $SE = 0.036$; incongruent: $M = 0.20$, $SE = 0.03$).

---

[22] It is important to realize that these scores reflect the patterns *after* participants were certain about their classification, because the "don't know" responses were treated as non-answers. In the first step it was found that the recognition speed was faster for the positive than for the negative emotions. Therefore, it is possible that the more positive classification in the first gate reflects the part of the population which is more certain about their answers, i.e. that an interaction is possible between the level of confidence and the extremity of the responses.

4.4.7 Summary

In this study, we used a *gating paradigm* to test the recognition speed for various emotional expressions from a speaker's face. Participants were presented with video clips of speakers who displayed positive or negative emotions, which were either congruent or incongruent. Using a gating paradigm, the clips were shown in successive segments which increase in length.

We first calculated the *confidence scores*, which are the number of times that the participants made a classification related to the number of times that they could not yet make a classification. We found the most NON-ANSWERS for the first gate, and the *congruent* emotions got more NON-ANSWERS than their incongruent counterparts. Further, in all conditions, the percentage of ANSWERS increased over the next gates, and reached a plateau after the fourth gate (640 *ms*). Also, the proportion of answers increased faster for the *positive* than for the negative emotions.

Next, we analyzed the *valence* of answers. Results show that participants are surprisingly accurate in their recognition of the various emotions, as they already reach high recognition scores in the first gate (after only 160 *ms*). Interestingly, this recognition plateau is reached earlier for *positive* than negative emotions. Finally, *incongruent* emotions get more extreme recognition scores than congruent emotions, and already after a short period of exposure, perhaps because the incongruent recordings contain more expressive displays.

Given the previous two perception experiments, the next section discusses an observational analysis which aims to find possible visual correlates of emotional expressions, both in the upper and lower area of the face.

4.5 Observational analysis

To gain further insight into which facial cues could have influenced the participants' categorization, we annotated all fragments in terms of a number of facial features. Although much is known about the prototypical expressions of emotions (Ekman, 1993), less is known about the difference in displayed *congruent* and incongruent facial cues for emotions (Wilting et al., 2006). Also, while past research has shown which facial cues are prototypical for pictures of emotions of joy and sadness, a second question is whether temporal dynamics such as the *duration* and the *intensity* of these cues can be successful in distinguishing between these positive and negative emotions, as these dynamics have already been shown to be successful in signaling the difference between congruent and incongruent displayals (Cohn & Schmidt, 2004; Valstar et al., 2006). Because temporal aspects of facial features are extremely difficult to assess manually and often require the use of advanced computer models (Cohn & Katz, 1998; Valstar et al., 2006), we chose to annotate solely whether or not a (number of chosen) feature(s) occurred, and the *subjective* intensity of these cues, rather than their exact duration and amplitude.

We concentrate on a small set of features. The chosen features are roughly comparable with Action Units described by Ekman and Friesen (1978), though there is not necessarily a one-to-one mapping to these Action Units. The choice of these features was based upon two restrictions: we wanted to score the *upper* as well as the *lower* face, and further we chose a set of features we assumed to reflect a *positive* as well as a *negative* emotion.

For the *upper* face we chose the following two features:

- RAISING THE BROWS. This feature resembles the Action Unit combination 1+2.
- FROWNING UPWARDS, i.e. raising the brows and frowning. This feature resembles the Action Unit combination 1+4.

For the *lower* face we chose the features:

- SMILING, i.e. pulling the corners of the mouth aside and up. This feature resembles the Action Unit 12.
- LOWERING THE MOUTH, i.e. pulling the corners of the mouth down. This feature resembles the Action Unit 15.

The labeling was performed by 3 coders, the first author of this chapter and two independent PhD-students, who were unfamiliar with the purpose of the current study, but who were experienced with visual annotations. The procedure was as follows. The coders watched the film fragments and labeled them using the set of 4 features. Each coder labeled each feature individually. The labeling process took place blind for condition. We asked the labelers to score the maximum intensity that the feature reached in the entire film clip. The presence of the feature was largely determined on the labelers' subjective impression of whether the feature occurred or not. Each feature was given a number between 0 and 2 to reflect different strengths, where 0 stands for a complete absence and 2 represents a very clear presence of the facial feature. The scores for the features were subsequently summed across the 3 coders resulting in an overall score between 0 and 6 for the respective features. For instance, when coder 1 scored a 2, and the other two coders scored a 1, the overall score was a 4. This way of computing of the intensity by summing up the scores of the three labelers is consistent with the method of Hirschberg, Litman & Swerts (2004) and Barkhuysen, Krahmer and Swerts (2004) to label auditory and visual degrees of hyperarticulation.

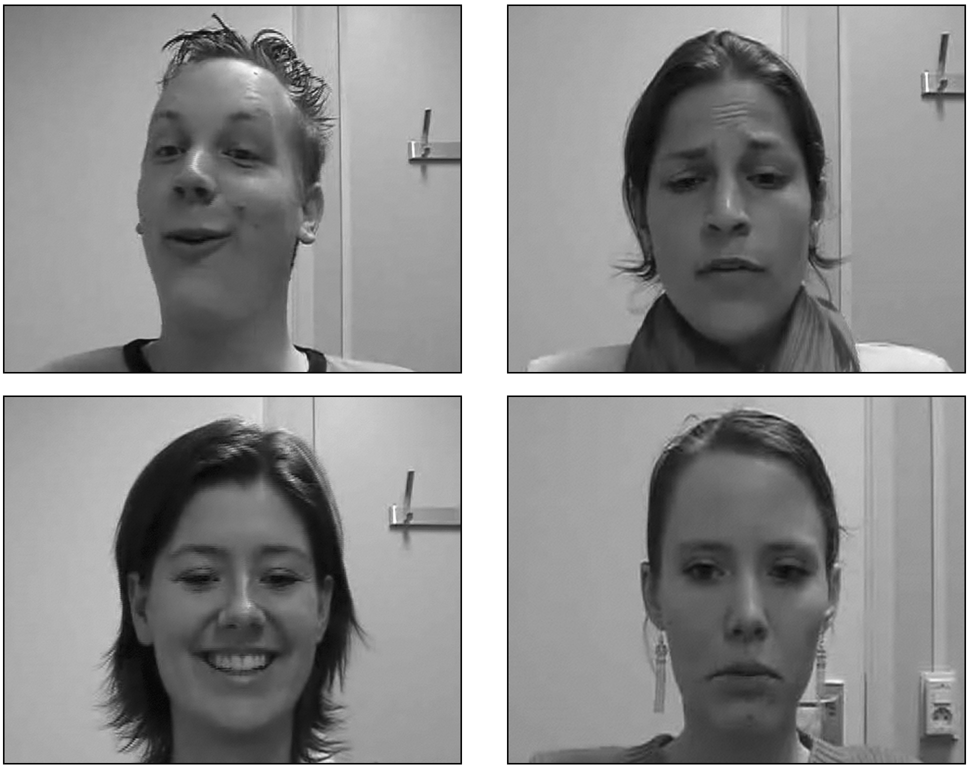

Figure **17** - Representative examples of the four annotated features: upper face (top) and lower face (bottom) expressions, with on the left hand side the positive and on the right hand side the negative versions

For each labeled feature, we computed the correlation between the three coders. The Pearson correlation was significant for all the 4 features (RAISING THE BROWS: r_{12} = 0.61, $p < .01$; r_{13} = 0.67, $p < .01$; r_{23} = 0.74, $p < .01$; FROWNING UPWARDS: r_{12} = 0.76, $p < .01$; r_{13} = 0.56, $p < .01$; r_{23} = 0.41, $p < .01$; and SMILING: r_{12} = 0.68, $p < .01$; r_{13} = 0.74, $p < .01$; r_{23} = 0.77, $p <$

.01). The correlation was somewhat lower for LOWERING THE MOUTH: r_{12} = 0.38, p < .01; r_{13} = 0.35, p < .01; r_{23} = 0.44, p < .01), but still significant.

4.5.1 Results

First of all, we present the general distribution of responses across the conditions in Table **19**.

Table **19** - Distribution of utterances from experiment 1 (standard errors between brackets) in terms of their mean scored intensity as a function of condition

Condition	*Inc. neg.*	*Negative*	*Neutral*	*Positive*	*Inc. pos.*	*Total*
Brows	2.20 (.61)	0.80 (.25)	0.80 (.59)	1.20 (.44)	2.70 (.86)	1.54 (.27)
Frowning	0.60 (.40)	0.30 (.15)	1.00 (.68)	0.70 (.47)	0.00 (.00)	0.52 (.19)
Smiling	0.40 (.40)	0.20 (.13)	0.40 (.16)	2.90 (.71)	3.70 (.54)	1.52 (.28)
Mouth	2.90 (.48)	2.10 (.43)	1.70 (.50)	1.30 (.40)	0.20 (.13)	1.64 (.22)

According to this table, the two features within either the upper (brows) or lower face (mouth) behave in an opposite way. Further, the intensity of the mouth is dependent upon condition, while the brows are independent from the valency of the condition.

Valency of the emotion of the speaker in the fragment

In this section, we explore to what extent there is a relation between the *valence* of the emotional state of the speaker in the fragment and the intensity of the annotated visual features described above. A univariate analysis of variance was performed for each of the separate features, with condition as independent factor (INCONGRUENT NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE, INCONGRUENT POSITIVE) and the feature as dependent factor (RAISING THE BROWS, FROWNING UPWARDS, SMILING and LOWERING THE MOUTH). There was a significant effect of *condition* on SMILING (F(4, 45) = 13.727, p < .001, η^2_p =.55), in the sense that the intensity of SMILING increases in the (congruent as well as incongruent) *positive* conditions (congruent: M = 2.9, SE = 0.446 and incongruent: M = 3.7, SE = 0.446). *Post hoc* analyses revealed that for SMILING, the *positive* conditions differ from the negative ones (p < .01), but the congruent conditions do not differ significantly from their incongruent counterparts (e.g. POSITIVE did not differ from INCONGRUENT POSITIVE). Further, the NEUTRAL condition differed from the positive ones (p < .01). There was also a significant effect of *condition* on LOWERING THE MOUTH (F(4, 45) = 5.940, p < .01, η^2_p =.346), in the sense that the intensity of lowering the mouth increases in the (congruent as well as incongruent) *negative* conditions (congruent: *M* = 2.1, *SE* = 0.410 and incongruent: *M* = 2.9, *SE* = 0.410). *Post hoc* analyses revealed that for LOWERING THE MOUTH only the INCONGRUENT POSITIVE condition differed from the two negative conditions. So, SMILING occurs more in the *positive* conditions, while LOWERING THE MOUTH occurs more often in the *negative* conditions (the latter only across congruent *and* incongruent conditions). This validates the data along with the well-known literature on facial expressions. The upper face did not vary consistently across conditions: the other two features were non-significant.

Incongruent vs. congruent emotions of the speaker in the fragment

Also, we are interested in whether there was a relationship between the intensity of these features and whether the speaker was displaying an emotional expression which was incongruent with the lexical content of the utterance. Although the univariate analysis of variance did not show an overall effect for raising the brows, inspection of Figure **17** and Table **19** tells us that the intensity of RAISING THE BROWS tends to increase in the *incongruent* conditions (negative: *M* = 2.2, *SE* = 0.586 and positive: *M* = 2.7, *SE* = 0.586), while the other three features do not seem to have a correlation. In order to test this further, we performed separate t-tests for each feature. In these tests, both incongruent conditions (negative and positive) as a group 'incongruent' were compared with a second group containing the two congruent conditions. It was shown that indeed only the feature RAISING THE BROWS was significant ($t = -2.529$, $df = 38$, $p < .05$). Therefore, the brows are raised more intense in the incongruent conditions.

Emotional intensity of each feature in the fragment as perceived by the judges

Next, we are interested in whether there is a relationship between the *intensity* of the annotated features for each fragment (as it was scored by the three coders) and the *perceived* emotional state of that fragment such as it was classified in experiment 1 (by the Czech judges). Figure **18** shows the mean intensity of each scored feature for each fragment as a function of the mean perceived emotional state (1 = very negative, 7 = very positive) in experiment 1 (in the VO condition). Again, the intensity of the mouth movements increases as the perceived valency of the emotional state grows stronger (in either direction), while the brows seem uncorrelated.

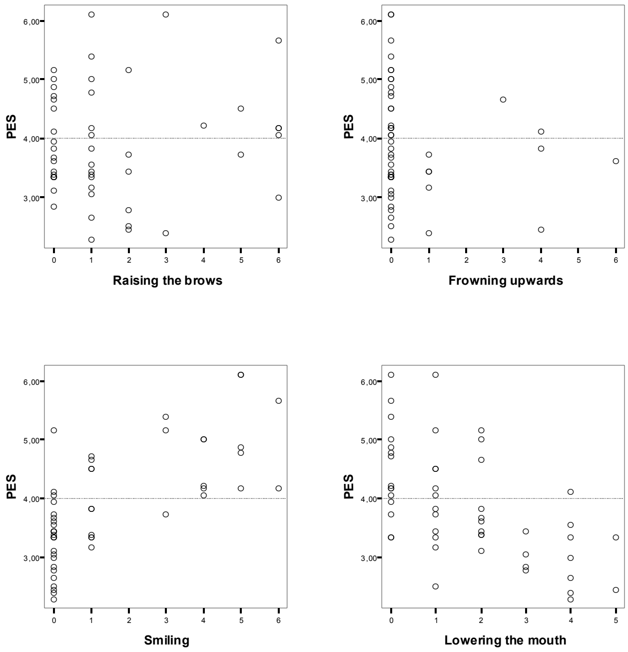

Figure **18** - The mean perceived emotional state (1 = very negative, 4 = neutral, 7 = very positive) such as each fragment was *classified* by the Czech participants in experiment 1, as a function of the mean intensity of each feature (0 = no intensity, 6 = very intense) for that fragment such as it was *scored* by the three coders

In order to test this, correlational analyses were performed between the 4 features and the mean perceived emotional state in experiment 1 (in the VO condition). The Pearson correlations for the features SMILING (r = 70.4, p < .01) and LOWERING THE MOUTH (r = -58.4, p < .01) were significant, though in opposite directions. The other two features were non-significant. Therefore, the more a fragment was perceived as positive, the more smiling occurred in the fragment. Vice versa, when the fragment was perceived as less positive, lowering the mouth was scored as more intense.

4.5.2 Summary

We were interested in the difference in occurrence of facial cues displayed in *positive* and negative conditions (both *congruent* and incongruent), and whether the intensity of facial cues can be useful for distinguishing between these conditions. The annotation analyses revealed that the occurrence of the features SMILING, LOWERING THE MOUTH and RAISING THE

BROWS varies consistently across conditions. The data showed that SMILING and LOWERING THE MOUTH correlated with the perceived emotion: SMILING is scored as more intense in the positive conditions, while LOWERING THE MOUTH is scored as more intense in the negative conditions. Also, because RAISING THE BROWS is scored as more intense in the incongruent conditions, RAISING THE BROWS can be used to detect whether a speaker is displaying an emotion which is opposite to the lexical content of the sentence. Another question was whether there is a relationship between the emotional state of the fragment as it was perceived in experiment 1, and the intensity of the annotated features as they were displayed in the fragments. The data showed that the more a fragment was perceived as positive, the higher the scored intensity of the SMILING was. Vice versa, when the fragment was perceived as less positive, LOWERING THE MOUTH was scored as more intense.

4.6 Discussion and conclusion

In this chapter, we investigated whether dynamic auditory and visual cues of emotional speech differ in perceptual strength, and how people deal with input coming from two modalities when they have to make judgments about a speaker's emotional state. In addition, we were interested in how fast people would recognize various emotions when presented with fragments of speech. Previous research has brought to light that listeners can successfully infer the emotional state of a speaker using information from a single modality (see e.g. Adolphs, 2002; Bachorowski, 1999; Banse & Scherer, 1996; Carroll & Russell, 1996; Scherer, 2003; Schmidt & Cohn, 2001). However, while there is much insight into how unimodal stimuli (either auditory or visual) are processed, less is known about the extent to which these modalities interact with each other. Also, while much research has been done in the field of audiovisual speech processing, less work has been done about cross-modal integration in the context of *emotional speech*. Next, there is more knowledge available about online auditory speech than about online visual speech, because many studies combined a dynamic speech signal with *static* facial images. In order to answer such research questions, we collected utterances in a semi-spontaneous way using a experimental paradigm eliciting *positive* and *negative* emotions. In this paradigm, the participants, while being videotaped, had to reproduce sentences increasing in emotional strength. The display of the negative or positive emotions could be congruent or incongruent with the lexical content of the sentences. Using these utterances, two perception experiments were carried out.

The first experiment was a classification experiment with Czech participants, to make sure that the participants could not rely on lexical cues. These participants were confronted with a selection of the recorded fragments, presented in three formats: AUDIOVISUAL (AV), VISION-ONLY (VO) and AUDIO-ONLY (AO). The task for participants was to indicate on a 7-point scale whether the speaker in the fragment was in a positive or a negative emotion. It was found that the highest scores were found in the (AV and) VO modalities, suggesting that the positive emotions are more clear in the visual modality, while the lowest scores were found in the (AV and) AO modality, suggesting that the negative emotions are more clear in the auditory modality. This is inline with other findings (Scherer, 2003, pp.235-236). Further, the AV modality was always scored best, suggesting that the combination of two modalities contains more information than a single modality, although the difference between the AV modality and the two single modalities was not significant. We also compared the classification of the Czech participants for the VO fragments with the results of the Dutch perception test with the same stimuli (Wilting et al., 2006), which lead to essentially the same results. Therefore, it seems that the recognition of emotions was not influenced by cultural differences (or by the fact that the Czech language may use different intonational patterns). See Elfenbein and Ambady (2003) for more discussion on such issues.

A second question we explored in this chapter is to what extent the recognition of emotion varies as a function of the *time* that people are exposed to the facial expressions of a speaker. In order to answer this question, a second experiment was conducted. In a gating experiment participants were offered with short parts of the original fragments increasing in length, from 160 *ms* (4 video frames) to 960 *ms* (24 video frames). After each gate participants had to indicate whether they believed that the speaker was in a positive or negative mood, or whether they could not make the distinction on the basis of the current gate. The results showed that the participants already reached high recognition scores in the first gate. The confidence of the participants, determined as the moment where they chose either a positive or a negative emotion rather than the neutral option, reached a plateau in the fourth gate. Interestingly, this recognition plateau is reached earlier for positive than negative emotions, which is comparable to the valency effects reported by Leppänen and Hietanen (2004). It is interesting to consider that in the latter experiment people need 635 *ms* processing time to correctly classify a picture of a happy face (95.5%), while in the current experiment 160-480 *ms* of information seems to be sufficient for classifying a film clip of a speaker in a positive state. As our *confidence scores* reach a plateau after 640 *ms*, which is consistent with the scores reported by Leppänen and Hietanen (2004), it might be useful to make a distinction between the capability of correctly classifying an emotion, which is already possible after only 160 ms, and the confidence a person has in his ability to make a correct classification, which reaches the top level only after 640 ms.

To ensure the ecological validity of the emotions studied, one has to consider several problems. A problem with many emotion studies is that they often rely on "acted" data. The work of Ekman (e.g. 1987, 1993), for instance, is based on posed photographs of actors, and also in speech research actors are frequently used. Also, the comparison of the role of different modalities is often investigated by using congruent versus incongruent speech analogous to McGurk tasks. This raised the question whether the incongruent emotions are representative of acted, voluntary emotions or whether they are representative of real, spontaneous emotions. Wilting et al. (2006) addressed this problem by creating an "acting" condition: by asking the participants to display an emotion which was opposite the lexical content of the sentences in the Velten task, such "incongruent" sentences become similar to "acted emotions" as speakers are displaying an emotion they are not feeling. The participants in the congruent task, on the other hand, were free to express the emotion invoked by the sentences. We can be sure that they were indeed feeling the congruent emotion because Wilting et al. (2006) tested which emotion they felt by presenting a survey afterwards. Although the survey indicated that the participant's emotions in the incongruent conditions was not different from the neutral condition, it would be interesting to further refine this test in the future, e.g. to find out whether there is indeed an absence of emotion or whether they may have started to feel a mixture of emotions. It would be nice if future studies could supplement the current study with findings of brain research or arousal measures such as galvanic skin response.

The first perception test showed that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, while the difference between incongruent and congruent speech is larger for the negative than for the positive emotions. This is in line with past research (Wilting et al., 2006), suggesting that incongruent emotions are perceived as more intense than congruent ones (possibly because they are displayed more intense). It is interesting to note, though, that especially the negative incongruent expressions appear to be "ironic", which may have been caused by the mismatch between the form and the lexical content (see e.g. Attardo, Eisterhold, Hay, & Poggi, 2003, for a discussion about multimodal markers of irony). It would be interesting to replicate the experiment in the future, where the participants have to utter a sentence containing a neutral lexical content after the last sentence of the (positive or negative) list, which may be used in the perception studies instead. Further, de Gelder and Vroomen (2000) report about the relative importance of the face above the voice for judging a (portrayed) emotion. Here, the difference between the two incongruent emotions was indeed somewhat larger in the VO modality than in the AO modality. Another interesting point is the difference between facial and vocal expression within the separate conditions. It seems that the Velten procedure did not elicit recognizable vocal expressions in the POSITIVE condition, whereas it elicited recognizable facial expressions. On the other hand, the speakers in the INCONGRUENT POSITIVE condition were able to display recognizable facial *and* vocal expressions. According to de Gelder and Vroomen (2000), there are differences in the effectiveness with which the face and the voice convey different emotions. The recognition of happiness, for example, remain accessible when the face is presented upside down, and also in focal brain damage patients where the recognition of several facial expressions is impaired. While in the voice, on the other hand, happiness is sometimes hard to tell apart from other emotions. Our results suggest that happiness can be detected in the voice when the speakers are *acting* that they are happy, while in fact, they do not necessarily feel that way.

The second perception test showed that the *incongruent* emotions received these more extreme recognition scores already after a short period of exposure. The gating results confirm earlier findings where incongruent emotions are perceived as more intense than congruent emotions (Wilting et al., 2006), as in the current experiment the former get more extreme recognition scores than the latter, and already after a short period of exposure, perhaps because the incongruent recordings contain more expressive displays. Horstmann (2002) reported that prototypical emotions resemble the most intense form of expressing an emotion. Perhaps when displaying an incongruent emotion, the speakers tend to display more prototypical expressions, in contrast to when they are free to express spontaneously whatever emotion they are feeling.

To gain further insight into which facial cues could have influenced the participants' categorization, we annotated all fragments in terms of a number of facial features. According to some models (Cohn & Schmidt, 2004; Valstar et al., 2006), dynamic facial expressions consist of an initial onset phase, a peak, and an offset phase. In the *onset phase* of an expression, the facial muscles contract until the facial expression reaches its apex. In the next phase, the facial expression is at its peak and does not change any further, until the start of the *offset phase*. Here, the facial muscles start to relax until the facial expression has returned to its neutral position (Valstar et al., 2006). The onset phase is usually very quick, ranging from 0.40 to 0.70 seconds in the case of smiles (Cohn & Schmidt, 2004). The participants' in our experiments needed only 160-480 *ms* for classifying a film clip of a speaker in a positive state, and their confidence scores reach a plateau after 640 *ms*, equaling the duration of an average onset phase. However, it is perfectly possible that displayed facial cues in the fragments were already at their apex, as we captured the speakers at the height of the induced emotion (by using only the last sentence of the list as a stimulus in the perception test).

We chose to annotate solely whether or not a (number of chosen) feature(s) occurred, and the *subjective intensity* of these cues, rather than their exact duration and amplitude. These features were RAISING THE BROWS, FROWNING UPWARDS, SMILING and LOWERING THE MOUTH. The occurrence of two other possible candidates, i.e. gaze and head movements, was too low, but these features seem to be correlated with end-of-utterance marking (Barkhuysen, Krahmer, & Swerts, 2008). We felt that the intensity of the scored features is a reflection of the displayed apex in the offered fragments. We investigated to what extent there is a relation between the valence of the emotional state of the speaker in the fragment and the annotated visual features described above, i.e. whether the *intensity* of facial cues can be successful in distinguishing between positive and negative emotions. It was shown that the intensity of the mouth was correlated with the intensity of the perceived emotion, in that when the mouth is lowered, the fragment is perceived as more negative, while the fragment is perceived as more positive when the mouth is smiling.

Further, we expected that the final intensity of the displayed cue can discriminate between congruent, "spontaneous", and incongruent, "acted" emotions, because posed smiles have a smaller amplitude (e.g. Carroll & Russell, 1997, in Cohn & Katz, 1998) and also the intensity of brow actions has been shown to be successful for distinguishing between spontaneous and posed expressions (Valstar et al., 2006), although it is not clear in what direction this relationship was. Our data showed that only raising the brows tends to increase in the incongruent conditions.

Next, we were interested in whether there is a relationship between the emotional state of the fragment as it was perceived in experiment 1, and the intensity of the annotated features as they were displayed in the fragments. The data showed that the more a fragment was perceived as positive, the more smiling occurred in the fragment. Vice versa, when the mouth was lowered more intense, the fragment was perceived as less positive.

Possibly, the *configuration* of features may be more important than simply distinguishing "which feature is responsible for what". Neurological research shows that faces are processed as a whole, apart from the full processing of individual features (Adolphs, 2002), and there are even more specialized routes for the processing of moving faces, i.e. dynamic, changeable configurations of facial features (Adolphs, 2002; Humphreys et al., 1993), although there are multiple interactions between the several pathways (Vuilleumier &

Pourtois, 2007). Also, the timing and coordination of the various regions of the face are usually off the mark in posed expressions (Ekman & Friesen, 1978). However, based upon the annotation results it is very likely that at least information from the *mouth* could have been very useful. Although the upper face in general, in particular the eyes, is reported as the most important source for emotion recognition, combining vocal expressions with facial expressions may draw attention to the mouth, unintentionally making the lower part of the face the most important source (de Gelder & Vroomen, 2000). It is therefore possible that in emotional *speech*, other facial features are important than in emotional expressions without speech.

Appendix

This appendix lists the Dutch variant of the Velten sentences used in this chapter.

Positive list

	Dutch	*English*
1.	Het is vandaag een dag als alle andere.	Today is a day like any other day.
2.	Toch voel ik me best wel goed vandaag.	Yet I feel rather good today.
3.	Deze dag zou wel eens een van mijn betere dagen kunnen zijn.	This day may be one of my better days.
4.	Als je een positieve houding hebt, gaat alles goed. Mijn houding is positief.	If your attitude is positive, everything will be allright. My attitude is positive.
5.	Ik heb energie en zelfvertrouwen in overvloed.	I have energy and self confidence in abundance.
6.	Ik voel me opgewekt en vrolijk.	I feel cheerful and gay.
7.	Ik denk dat vandaag alles verder heel goed zal gaan.	I think that today everything will go very well.
8.	Mijn mening over de meeste zaken is weloverwogen.	My opinion about most matters is well-considered.
9.	Ik zit zo vol energie; ik kan heel lang doorgaan zonder te slapen.	I am so full of energy; I can continue for a long time without sleeping.
10.	Ik kan dingen vandaag goed inschatten; niemand kan mij van gedachten doen veranderen.	I can assess things good today; nobody can change my mind.
11.	Als ik wil, kan ik er voor zorgen dat alles goed gaat.	If I want to, I can make sure that everything goes allright.
12.	Ik ben enthousiast en voel me zelfverzekerd.	I am enthusiastic and feel self-confident.
13.	Ik denk dat er mooie tijden aankomen.	I think that there will be good times.
14.	Ik heb zin om met iedereen te praten.	I like to talk to anybody.
15.	Ik weet heel goed dat ik mijn doelen kan bereiken.	I know very well that I can achieve my goals.
16.	Ik voel me sterk en vitaal.	I feel strong and fit.
17.	Ik voel me zo levendig en krachtig vandaag. Ik kan alles aan.	I feel so lively and powerful today. I can cope with anything.
18.	Niemand kan me stoppen vandaag!	Nobody can stop me today!

19.	Voortaan zal ik zogenaamde "problemen" niet meer groter maken dan ze zijn.	In the future I won't blow up so-called "problems".
20.	Ik heb geen tijd om me zorgen te maken; ik ben veel te druk bezig met andere dingen.	I don't have time to worry; I am too busy with other things.
21.	Ik voel me verbazingwekkend goed vandaag!	I feel surprisingly well today!
22.	Ik voel me creatief en inventief vandaag.	Today I feel creative and inventive.
23.	Ik voel me super!	I feel super!
24.	Alles ziet er goed uit; alles ziet er geweldig uit!	Everything seems ok; everything looks great!
25.	Ik zie alles van de zonnige kant.	I see everything from the sunny side.
26.	Ik voel me erg opgewekt en levendig.	I feel very cheerful and lively.
27.	Ik zie alles scherp en in een nieuw daglicht.	I see everything sharp and in a new daylight.
28.	Mijn geheugen werkt voortreffelijk vandaag.	Today my memory works excellent.
29.	In een goede stemming als deze werk ik snel en lukt alles meteen.	In a good mood like this I work quickly and everything works out immediately.
30.	Ik kan me goed concentreren op alles wat ik doe.	I can concentrate well upon everything I do.
31.	Ik denk helder en snel.	I think sharp and quickly.
32.	Het leven is zo leuk; het geeft me zoveel voldoening.	Life is so enjoyable; it gives me so much pleasure.
33.	Alles zal vandaag steeds beter gaan.	Everything will go better again and again.
34.	Ik voel me energiek. Ik wil iets doen!	I feel energetic. I want to do something!
35.	Ik heb alles onder controle.	I have everything under control.
36.	Ik zou wel goede harde muziek willen horen!	I would like to hear good hard music!
37.	Dit is geweldig; ik voel me echt goed.	This is great; I feel really great.
38.	Dit is zo'n dag waarop ik ervoor ga!	This is one of those days that I go for it!
39.	Ik zit vol energie.	I am full of energy.
40.	God, wat voel ik me geweldig!	God, I feel great!

Negative list

	Dutch	*English*
1.	Het is vandaag een dag als alle andere.	Today is a day like any other day.
2.	Toch voel ik me neerslachtig vandaag.	Yet I feel depressed today.
3.	Ik voel me best sloom op het moment.	I feel rather slow at the moment.
4.	Soms voel ik me zo moe en somber dat ik alleen maar wil zitten.	Sometimes I am so tired and gloomy that I just want to sit down.
5.	Het lijkt wel alsof iedereen energie heeft, behalve ik.	It seems that everyone has energy, except me.
6.	Mensen irriteren me. Waarom laten ze me niet met rust?	People irritate me. Why don't they leave me alone?
7.	Ik heb het gevoel dat ik nauwelijks vooruit kom.	I feel like I hardly proceed.
8.	Soms voel ik me zwak en verward en loopt alles wat ik doe in de soep.	Sometimes I feel weak and confused and everything I do smashes up.
9.	Van een beetje inspanning word ik al moe.	I get tired from a little effort.
10.	Ik voel me vandaag verschrikkelijk moe en alles kan me gestolen worden.	Today I feel terribly tired and I don't care about anything.
11.	Ik ben kapot. Mijn lichaam voelt uitgeblust en zwaar aan.	I am broken. My body feels washed out and heavy.
12.	Ik begin me slaperig te voelen. Ik dwaal steeds af.	I start to feel sleepy. I stray off all the time.
13.	Mijn leven is zo vervelend. Elke dag diezelfde sleur is deprimerend.	My life is so annoying. Everyday that same routine is depressing.
14.	Ik kan me dingen nu niet zo goed herinneren.	I can't remember things very well right now.
15.	Ik kan gewoon geen knopen doorhakken. Het is zo moeilijk om simpele beslissingen te nemen.	I just can't make decisions. It is so hard to take small decisions.
16.	Ik heb zin om mijn ogen dicht te doen en hier ter plekke te gaan slapen.	I feel like closing my eyes and to sleep right here.
17.	Ik ben niet erg kwiek; ik voel me lusteloos en verdrietig.	I am not very spry; I feel listless and sad.
18.	Ik ben geen stuiver waard.	I am worthless.
19.	Ik voel me belabberd. Mijn gezondheid is niet zoals het zijn moet.	I feel terrible. My health is not what it should be like.
20.	Niemand begrijpt me als ik klaag of me ongelukkig voel over mezelf.	Nobody understands me when I complaint or feel unhappy about myself.
21.	Ik ben onzeker over mijn toekomst.	I am insecure about my future.

22.	Ik ben moedeloos en ongelukkig met mezelf.	I am despondent and unhappy with myself.
23.	Alles is nu slechter dan toen ik jonger was.	Everything is worse than when I was younger.
24.	Zoals ik me nu voel, ziet de toekomst er saai en hopeloos uit.	The way I feel now, the future looks boring and hopeless.
25.	Ik vind het ontzettend moeilijk om belangrijke beslissingen te maken.	It is very hard to take important decisions.
26.	Ik voel me moe en depressief; ik heb geen zin om iets te doen.	I feel tired and depressed; I don't want to do anything.
27.	Alles gaat makkelijker en beter bij andere mensen dan bij mij.	Everything goes easier and better in other people than in me.
28.	Vaak maken mensen me erg boos. Ik ben liever alleen.	Often people make me very angry. I prefer to be alone.
29.	Ik kan niet goed over mijn problemen praten met anderen.	I can't talk about my problems very well with other people.
30.	Mensen luisteren nooit echt naar me.	People never really listen to me.
31.	Ik heb me weleens zo alleen gevoeld, dat ik had kunnen huilen.	Sometimes I felt so alone than I could have cried.
32.	Soms wou ik dat ik dood was.	Sometimes I wish I was dead.
33.	Mijn gedachten zijn zo traag en somber; ik wil niet denken en niet praten.	My thoughts are so slow and gloomy; I don't want to think or to talk.
34.	Ik geef nergens meer om. Het leven is gewoon niet leuk.	I don't care about anything anymore. Life is just not enjoyable.
35.	Ik ben zo moe.	I am so tired.
36.	Ik heb veel te slechte dingen meegemaakt in mijn leven.	I have experienced too many bad things in my life.
37.	Alles is waardeloos en leeg.	Everything is worthless and empty.
38.	Ik heb geen zin om iets te doen.	I don't want to do anything.
39.	Alle tegenslagen in mijn leven achtervolgen me.	All misfortunes in my life haunt me.
40.	Ik wil slapen en nooit meer wakker worden.	I want to go to sleep and never wake up again.

5 General discussion

The three studies described in this thesis were intended to improve the understanding of the production and perception of audiovisual speech. In this chapter we summarize and discuss the main findings. We will also suggest directions for future research.

5.1 Summary and conclusions

In this thesis we investigated speech from a multimodal perspective. In the first chapter we discussed which starting points form the basis of our research. We started with the assumption that speech is multimodal in its nature, that the role of the sender as well as of the receiver is important, and that we should investigate natural interactions.

These assumptions had important implications for our general research methodology. First, we recorded the audiovisual speech in order to be able to present it in different modalities later on. These modalities were a bimodal, audiovisual condition (i.e. the original recording), or a unimodal condition, which was either vision-only (only the visual signal, i.e. facial expressions, stored) or audio-only (only the auditory signal, i.e. the voice, stored). Second, we elicited the audiovisual speech under controlled circumstances, such that we could use the audiovisual speech as stimulus material for perception tests. Third, we had the participants engaging in a natural interaction. These assumptions thus determined the design of the experiments.

Further, the three studies reported in this thesis focused on different components of an interaction. These components are (1) how dialogue participants provide feedback, e.g. how they signal and detect communication problems, (2) how participants regulate turn-taking, e.g. how they display and detect end-of-utterance marking, and (3) how participants display and perceive emotions.

In the rest of this section we will discuss what the main findings were in these three studies. In section 5.2, we will discuss these findings in the light of the starting points.

5.1.1 The detection of communication problems

In *Chapter 2* we showed that participants are able to detect whether there is a communication problem in a dialogue, on the basis of only short fragments. In a series of three perception tests participants were offered film fragments (without any dialogue context) of speakers interacting with a spoken dialogue system, offering a train travel advice. In half of these fragments, the speaker is or becomes aware of a communication problem. Participants had to determine by forced choice which are the problematic fragments. The perception tests reflected different phases in the dialogue: verification questions of the system (e.g. "*So you want to travel to Amsterdam?*"), negations ("*no*") and slot fillers (e.g. "To which station do you want to travel?" - "*Utrecht*") on the part of the speaker. In all three tests, participants were capable of performing this task to some extent, but with varying levels of correct classifications. The negations were most difficult, the verification questions were classified best. Because the verification questions showed people listening to a system's question, it could be stated that participants perform better in the 'vision-only' condition (i.e. verification questions) than in the bimodal conditions (i.e. negations and slot-fillers).

Because the negations were the shortest stimuli, we speculated that in longer stimuli there may be more cues available. Also, the classifications differed across speakers, which led us to the question whether it is possible that different speakers show different cues. On the basis of these two questions we decided to do an additional observational analysis, in which we showed that more problematic contexts lead to more dynamic facial expressions, in line with earlier claims that communication errors lead to marked speaker behavior. Both hyperarticulation (i.e. exaggerated speech, e.g. "AM....ste:rr..dam") and the amount of visual variation (which is the sum of several individual facial features) played a role. Also, the presence of these features influenced the perception of problems. We concluded that visual information from a user's face is potentially beneficial for problem detection, and that a system could use this information by automatic facial tracking, in order to monitor the user's frustration and the concurrent presence of errors.

5.1.2 The detection of end-of-utterance

In *Chapter 3* we showed that speakers also employ visual cues, apart from auditory cues such as intonation, rhythm, and pausing, to indicate that they are at the end of their utterance. Speaker utterances were collected via a novel semi-controlled production experiment, in which participants provided lists of words in an interview setting (e.g. "What are the colors of the Dutch flag?" - "*Red.. white.. blue*", or "What are the odd numbers between ten and zero?" - "*Nine.. seven.. five.. three.. one*"). The data thus collected were used in two perception experiments, which systematically compared responses to unimodal (audio-only and vision-only) and bimodal (audiovisual) stimuli.

The first experiment was a reaction time experiment, which revealed that participants are significantly quicker in end-of-utterance detection when confronted with bimodal or audio-only stimuli, than for vision-only stimuli. Also, short stimuli (e.g. "*seven*") led to longer reaction times than long stimuli (e.g. "*nine.. seven*"). Because no significant differences in reaction times were found between the bimodal and audio-only condition, a second experiment was conducted.

The second experiment was a classification experiment, and showed that participants perform significantly better in the bimodal condition than in the two unimodal ones. Also, short stimuli were more difficult to classify than long stimuli. Further, non-final stimuli (e.g. "*nine.. seven*" versus "three.. one") were classified better, but there was a small response bias. Both the first and the second experiment revealed interesting differences between speakers in the various conditions, which indicates that some speakers are more expressive in the visual and others in the auditory modality.

We conducted an additional observational analysis, in which we showed that several auditory as well as visual cues seemed to play a role in the judgment of finality. Further, there were large differences between speakers in the amount and type of features displayed, also depending on the place in the utterance. It was suggested that many of these features may be rhythmically distributed over the different phases within an utterance (e.g. a rhythmic pattern of nodding on words, or diverting and returning the head).

5.1.3 The detection of emotions

In *Chapter 4* we showed that people classify emotions in a different way across different modalities, depending on the valence of the emotion (i.e. positive or negative) and whether the speaker is experiencing an emotion congruent with the lexical content. Both experiments were based on tests with video clips of emotional utterances collected via a variant of the well-known Velten method. More specifically, we recorded speakers who displayed positive or negative emotions, which were congruent or incongruent with the (emotional) lexical content of the uttered sentence (e.g. "God, I feel great!" uttered in a 'happy' or 'unhappy' way).

The first experiment is a perception experiment in which Czech participants, who did not speak Dutch, had to rate the perceived emotional state of Dutch speakers, and showed no overall differences between the modalities. It was found that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, where the difference between congruent and incongruent emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between congruent and incongruent emotions were found for the audio-only condition, which suggests that displaying an incongruent emotion has a particularly strong effect on the spoken realization of emotions.

The second experiment used a gating paradigm to test the recognition speed for various emotional expressions from a speaker's face. In this experiment participants were presented with the same clips as experiment 1, but this time presented vision-only. The clips were shown in successive segments (gates) which increase in length (e.g. from "Go.." to "God I feel gr.."). Results show that participants are surprisingly accurate in their recognition of the various emotions, as they already reach high recognition scores in the first gate (after only 160 milliseconds). Interestingly, the recognition scores rise faster for positive than negative conditions. Finally, the gating results suggest that incongruent emotions are perceived as more intense than congruent emotions, as the former get more extreme recognition scores than the latter, already after a short period of exposure.

We conducted an additional observational analysis, in which we showed that the occurrence of three visual features differed depending on the valence of the emotion and whether the speaker is experiencing an emotion congruent with the lexical content. It was also shown that the occurrence of these features was related to the perceived emotional state.

5.2 Discussion

On the basis of the assumptions we formulated several research questions, which were addressed in the three studies. These research questions covered the role of the modalities, the role of the sender and the receiver, and the role of natural data. In *Chapter 1* we formulated some questions for each domain. We will now discuss the results reported above in the light of these questions.

5.2.1 The role of the modality

The first question was *whether 'visual' speech is informative, so that it may express a 'visual' equivalent of prosody*. In *Chapter 2*, participants were able to classify whether there was a communication problem on the basis of fragments where a sender was only listening to a system's question. In the classification experiment described in *Chapter 3*, the vision-only condition yielded around 75% correct classifications of finality. In *Chapter 4*, the classification of emotions was not worse, i.e. different, for the vision-only condition than for the other conditions. All three chapters thus suggest that it is possible to interpret signals from the visual modality only. Is it thus likely that 'visual' prosody does exist? The current results suggest that the visual modality is actually being used in the perception of speech recorded in different circumstances, reflecting user frustration, emotions, and different phases of an utterance. When prosody is defined in a broad way, including the expression of attitudes and emotions, it is likely that the visual modality can be used for the expression of a variety of prosodic features and functions. Future research will have to specify what the exact nature and role of such visual information will be.

Second, we were interested *whether speech coming from different modalities is integrated by the receiver, i.e. whether the modalities complement or obstruct each other*. In general, the combination of two modalities provided better results than one. In *Chapter 2*, participants can correctly classify communication problems in the 'bimodal' stimuli. Although the stimuli that were 'vision-only', i.e. participants were listening to a system's question, scored better than the 'bimodal' stimuli, i.e. the negations and the slot fillers, this was possibly just because the verification stimuli were longer and contained therefore more information for a correct classification. In *Chapter 3*, the classification of the finality of a fragment was better for the bimodal condition than for the unimodal conditions, and the bimodal condition elicited also faster reaction times. There were some interaction effects suggesting that the information value of a modality may depend upon which cues were displayed by a speaker, and upon the length and the type of the fragment. In *Chapter 4*, the classification of emotions was not different for the three conditions, but there were some interaction effects depending upon the congruency (incongruent, 'acted' or congruent, 'real') and the valence of the emotion (i.e. positive or negative).

These results suggest that integration occurs to some extent. For detecting end-of-utterance a combination of two modalities indeed provides more information than each single modality. Further, the information value of a modality depends on other factors, such as the speaker involved, characteristics of the stimuli (e.g. the length, the valence, or the (final or non-final) location within the original utterance), and the role that the utterance has within the dialogue (e.g. a system's question, an answer of the participant).

The third question was *which modality is the most important.* In *Chapter 2*, the 'visual' stimuli were classified better than the bimodal stimuli, though, as discussed above, this may have been an experimental artifact. However, the observational analysis showed that in the three types of stimuli, the auditory cues are stronger correlated with the perception of problems than the visual cues alone, as well as with the presence of problems. In *Chapter 3*, detecting end-of-utterance in the visual condition elicited longer reaction times than in the auditory condition, but both unimodal conditions elicited around the same percentage of correct classifications in the second experiment. The observational analysis showed that both auditory and visual cues were related to the finality of the fragments, and may therefore be equally useful for end-of utterance detection. In *Chapter 4*, both unimodal conditions scored the same results.

These results suggest that the auditory and the visual modality are equally informative, but that they can complement each other: when information is missing in one modality the participants can turn towards the other modality. It is possible that the role of the modality may be differ across prosodic subdomains, such that the visual modality may be more important in some domains while the auditory modality may be more important in others. However, the studies do not directly support this, as both in the domain of end-of-utterance marking and in the domain of emotions, there are no clear differences between the unimodal conditions.

5.2.2 The role of the sender and the receiver

Within this domain, the first question was *which cues senders actually display in audiovisual speech and to which of these available cues observers are sensitive.* In *Chapter 2* it was shown that senders displayed auditory cues such as hyperarticulation, as well as a number of facial features captured under the amount of visual variation, to signal the presence of problems. Both played a role in the perception, but hyperarticulation was more important. The visual variation consisted of visual cues such as frowning, repeated head gestures, eyebrow raising, smiling, eye movements, mouth opening and diverted head position. In *Chapter 3*, it was shown that low- and high-ending pitch, creakiness of the voice, returning eyes to the initial position, returning head and posture to the initial position, closing the mouth, blinking, nodding, and lowering the brows to the initial position all were displayed by senders and could potentially play a role in the judgment of finality. We did not test to which of these cues observers paid attention. In *Chapter 4*, we annotated only the visual cues that were displayed by senders, such as raising the brows, smiling, frowning upwards, and lowering the mouth. Also, raising the brows seemed to have a correlation with displaying an incongruent emotion, and may have caused the more extreme perception of these emotions. It was shown that the visual features smiling and lowering the mouth were correlated with the perception of positive versus negative emotions respectively.

It can be concluded that senders display a wide spectrum of behaviors. They display auditory and visual cues, large movements such as posture changes as well as subtle movements such as eye movements. We only annotated *what* features were displayed and did not further divide them into the *type* of behavior. However, visual inspection of the data showed us that senders employed several different types of behavior. Senders displayed global movements, i.e. movements that are spread over the whole utterance, such as raising the brows and lowering them again, as well as local behaviors, such as nodding on an individual word. Further, the use of a particular behavior highly depends on the context, e.g. raising the brows can indicate a communication problem, the start of a new utterance, or is used when displaying an incongruent emotion, perhaps related to raising the brows when marking sentence or word stress. Also, behaviors seemed to be combined in a rhythmic pattern, e.g. in *Chapter 3*, change of posture was often accompanied with change of head and eye direction, and during the utterance there seemed to be a rhythmic pattern of nodding on individual words, in combination with changes in eye direction. In future research it would be interesting to systematically test all these differences: between global and local cues, the role of the direct context, and the timing of the cues. Not only are there different types of cues, but senders can also combine these types within a single utterance. For example, in *Chapter 2*, when observers display that there are communication problems, this is not only because it has a function in the dialogue, but also because presumably they may have started to feel a mixture of emotions: they have become frustrated. In this light it is plausible that senders not only signal feedback cues, but also started to display cues that signal (a negative) emotion.

Second, we wanted to explore *whether there are individual differences between senders.* In *Chapter 2*, some senders displayed more cues than others, and also different cues. Further, senders displayed different cues depending on the role they had in the dialog: when they were listening to the system (i.e. verification questions) they displayed other cues, than when they were answering a question. In *Chapter 3*, again some senders displayed more cues than others, and also different cues. Some senders displayed more visual cues, other more auditory cues. Further, some senders seemed to have their individual style, e.g. one sender did often not return the gaze but instead nodded on the final position. In *Chapter 4*, we did not test for sender differences. In sum, there are clear sender differences in their preferences for a modality, in which cues they show and in how many, and in the choice for one cue over another to serve the same function (individual style). The distribution of cues over the auditory versus visual modality seemed to be sender-dependent. We did not further explore these sender differences, but they may have been caused by factors as sex, age, personality, and culture (see e.g. Cohn et al., 2002; Matsumoto, 1990).

5.2.3 The role of natural data

Here, we wanted to know *how we can use natural data in an experimental design, so that we can generalize our results to real interactions.* In *Chapter 2*, the elicited dialogues were very natural. The train-travel advice system was a fully operational system, and there were no restrictions on how the recorded participants had to solve the tasks. Also, the experimenter had left the room in order no to affect their behavior. Because the structure of a human-machine dialogue has limited degrees of freedom, it was easy to use the recordings in an experimental set-up. In *Chapter 3*, the participants engaged in an interview situation, and were not aware of the fact that their elicited answers were necessarily in the form of a list due to the nature of the questions. This ensured that the answers were natural and structured at the same time. In *Chapter 4*, we used an elicitation paradigm that has been shown to successfully invoke emotions (Velten, 1968). The interaction in this chapter was perhaps not as natural as in the other two, as the structure and lexical content of the sentences was given. Yet it is likely that the Velten paradigm may have successfully elicited emotions that were actually felt by the speakers, at least in the congruent conditions. In half of the cases, the speakers had to display an emotion that was inconsistent with the valence of the uttered sentence. This enabled us to compare 'natural' emotional expressions with 'artificial' emotional expressions.

Although the evoked expressions were probably natural, it is not completely clear to what extent we can generalize the results to daily life situations. In *Chapter 3* and *4*, the camera was clearly visible. Based upon incidental remarks of the participants, they were aware of this presence, and this awareness may have been a factor that affected their behavior. There seems to be an *observer's paradox*, i.e. in order to record a natural dialogue we had to use a camera, but the presence of this very camera could have disrupted a natural conversation (see Cieri, Miller, & Walker, 2002). After a while though, participants sometimes seem to forget to actively monitor their speech (Cieri et al., 2002). In *Chapter 2*, participants were led to believe that the camera was part of a videophone system, so instead of fulfilling the role of an external observing device, the participants may have regarded the camera as a true dialogue participant whom they were having a conversation with. They may thus have been less aware of the fact that it was also recording them.

An alternative would have been not to notify the participants that they were filmed at all, by using a hidden camera. However, observing participants with a hidden camera is difficult and perhaps unethical. The problems created by observing audiovisual behavior with a camera are not unique to this thesis, but are a general problem in audiovisual speech research, and we believe that the use of a visible camera is a good choice, as long as there is no better alternative.

5.3 Future directions

In this section we will discuss topics that we did not address in this thesis, but which may be fruitful directions for future research.

5.3.1 Extending the context

Audiovisual expressions are always displayed in a context. A first type of context is the dialogue in which the expressions are embedded. This may play a role when the emotions suggested by the context are incongruent with the emotions suggested by the facial expression (Fridlund, 1994, pp.237-238). In *Chapter 3*, we addressed the difference between local parameters (covering individual words or speech segments) and global parameters (stretching over one or more utterances) in detecting the end of an utterance. This type of context may also be important in other situations, for example in human-machine interaction. Indeed, a study showed that prosodic features are highly dependent upon the local context, where in the case of problem detection this is the most recently asked system question type (Lendvai et al., 2002). This study states that when separate classifiers are trained on subsets of the data that are split by the local context, the learners profit much more from prosodic information.

Another type of context is the social and/or cultural context. The facial display of emotions, for example, is strongly influenced by the social-cultural context, as well as the interpretation of these expressions (Matsumoto, 1990; Matsumoto & Ekman, 1989). An individual's behavior is often regulated by *display rules*, which can vary across cultures (Ekman & Friesen, 1975, p.24, pp.137-140, pp.154-155; Matsumoto, 1990). Not only the mere presence of other people is important, but also the *role* of the other in a social context (Jakobs, 1998). The vocal expression of emotions and the corresponding emotional speech may vary across cultures and social contexts as well. In this thesis, we tried to eliminate the role of this context as much as possible by the use of an experimental design. In *Chapter 4* though, we used observers from different language groups. However, as we were not interested in cultural differences in perception, we systematically compared the findings of the Czech observers with those of the Dutch observers. Because the findings were not different, it is safe to conclude that the cultural factor was not an important factor in this perception test. Yet, it may be very interesting to vary these two types of contexts in future research.

5.3.2 Extending the human factor

In this thesis we used a group of observers to judge audiovisual expressions out of the original context. By doing so, we assumed that these observers formed a homogenous group, and shared the same perspective as they would have had when they had been active participants in the recorded interactions. These assumptions may be challenged in additional research.

Although we addressed individual differences in the sender in *Chapter 3* and *4*, individual differences of the receiver may also be important. Kita and Özyürek (2003) describe a model in which cognitive plans for speech and gestures are based on linguistic and spatial representations, and in which these two representations co-evolve during the production of (audiovisual) speech. It would be interesting to measure in a pretest whether individuals differ in their preference for a language-oriented or for a visual-spatial thinking style, and if so, to subsequently test whether this preference correlates with the preference for one modality over another in processing audiovisual speech, i.e. a preference to process facial gestures rather than the auditory speech signal.

Other theories state that prosody itself is lateralized, in that the left hemisphere appears to be relatively dominant in perceiving linguistic prosody, and the right hemisphere in perceiving emotional prosody (Ross et al., 1988, Baum and Pell, 1999, and Pell, 2002, in Wilson & Wharton, 2006). It would be interesting to test whether an individual's preference for a (perhaps hemisphere-based) thinking-style correlates with the preference for processing linguistic versus emotional prosody, and therefore, whether these two groups of observers may vary over the type of prosody under investigation. For example, in *Chapter 3* there may have been more linguistic cues available in the material, and in *Chapter 4* the recordings may have contained more emotional cues.

A second assumption was the perspective of the perceiver of the audiovisual material. Perceivers can take the role of addressees and overhearers (Schober & Clark, 1989). Speakers actively try to make themselves understood by addressees, who in turn display back-channel signals. This gives addressees an advantage over people who are listening by the conversation but are not actively participating, i.e. the *overhearers*. By displaying back-channel signals, the addressees let the speakers fill in the gaps in their understanding with the information they are still lacking. As our observers did not share the common ground in the conversation (also because we eliminated the context before presenting the recordings, i.e. the utterance in which the stimuli were embedded), and the speakers may have behaved differently because they knew that there are overhearers 'present' (for example, because there was a camera), there is not necessarily a one-to-one mapping from our observers to addressees in daily life situations. The fact that the observers could successfully classify the situation from where the fragments were cut as "a communication problem", or "not at the end of an utterance", or "the speaker is in a positive emotion", suggests that this was not a large problem, but it may be interesting to insert this as an extra factor in future designs (e.g. by comparing online perception - where the observers are active participants as well - with off-line perception).

5.3.3 Extending the level of detail

It is desirable to pursue a further refinement of the *cues* in the audiovisual material. The studies described in this thesis were investigated with perception tests. While this will tell us something about the perception of audiovisual speech, it needs to be further refined in a later stage. The problem with perception tests is that it is difficult to know whether a receiver will use all cues just because they are there, that receivers only use a selection of all available cues, or that the use of a particular cue is dependent on the modality. Specific hypotheses following from the results presented in this thesis can be tested in follow-up research by the use of artificial simulation such as Embodied Conversational Agents (see e.g. Krahmer & Swerts, 2004), or by more detailed cue measurement techniques. A suitable method is the use of eye tracking techniques (Scherer, 2003). Another interesting possibility is to use the recordings from the perception tests as input to an audiovisual speech recognizer, which pays attention to several features, such as head motion, and audiovisual motion in general, in the implemented algorithms (see e.g. Wang, Demirdjian, & Darrell, 2007). With this technique it can be established which features are the most successful to cross an established recognition threshold[23]. These methods are not implemented in this thesis, but they represent an interesting area for future research.

We also propose a more detailed measurement of the *timing* involved in the perception of audiovisual material. The event-related brain potential (ERP) technique - which is based upon electroencephalography (EEG) - may be useful in investigating the dynamics of prosodic processes (see e.g. Steinhauer et al., 1999). ERP is very suitable for the *online*

[23] Note that although a feature can be useful for a succesful recognition by a machine, this does not necessarily mean that a human being uses the same feature for accomplishing a succesful recognition.

monitoring of speech because in case of an auditory signals, it is difficult to tell what part of the signal a listeners pays attention to due to the left-to-right nature (and therefore: incremental processing) of speech (see e.g. Marslen-Wilson and Tyler, 1980, in Grosjean, 1983; Rietveld & van Heuven, 2001, pp.294-295). In section 3.6, we have speculated that it is possible that a faster perception in the AUDIOVISUAL condition (as opposed to the unimodal conditions, i.e. the AUDIO-ONLY and the VISION-ONLY condition) could be caused by the fact that the different modalities might have served as sources providing complementary information, thus resolving potentially ambiguous information (compare Kim et al., 2004; Schwartz et al., 2004). Does the combination of two modalities invoke ambiguous conflicts or, on the contrary, can it solve ambiguities? It would be interesting to map the trajectory from where an ambiguity starts to where it is solved[24].

[24] ERP-responses have already been identified that can signal ambiguities. Researchers have shown that there is a positive deflection (P2), which peaks higher for congruous than for incongruous (emotional) stimuli and may be a signal of cross-modal integration (Balconi & Carrera, 2007; and Pourtois et al., 2000, for an other cross-modal integration related deflection), as well as a negative deflection (N2), which may be a signal of the type of emotion (Balconi & Carrera, 2007). These time-related results could be combined with structural imaging techniques that cover brain areas involved in processing auditory prosody (see e.g. Ross, 2000), or emotional facial expressions (see e.g. de Gelder, Vroomen, Pourtois, & Weiskrantz, 2000).

Bibliography

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Review, 1*(1), 21– 61.

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., & Öhman, T. (1998). *Synthetic faces as a lipreading support.* Paper presented at the 5th International Conference on Spoken Language Processing (ICSLP '98), Sydney, Australia.

Ahrenberg, L., Jönsson, A., & Thurée, Å. (1993). *Customizing interaction for natural language interfaces.* Paper presented at the Workshop on Pragmatics in Dialogue, the 14th Scandinavian Conference of Linguistics & the 8th Conference of Nordic and General Linguistics, Göteborg, Sweden.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge (GB): Cambridge University Press.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research, 16*(2), 243-260.

Aubergé, V., & Cathiard, M.-A. (2003). Can we hear the prosody of smile? *Speech Communication, 40*(1-2), 87-97.

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

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*: Cambridge University Press.

Balconi, M., & Carrera, A. (2007). Cross-modal perception (face and voice) in emotions: ERPs and behavioural measures. *Neuropsychological Trends, 1*, 43-64.

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

Barkhuysen, P. N., Krahmer, E. J., & Swerts, M. G. J. (2004). *Audiovisual perception of communication problems.* Paper presented at the 2nd Speech Prosody Conference (SP '04), Nara, Japan.

Barkhuysen, P. N., Krahmer, E. J., & Swerts, M. G. J. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America, 123*(1), 354-365.

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often? *Nature, 300*(5894), 744-747.

Benoît, C., Martin, J.-C., Pelachaud, C., Schomaker, L., & Suhm, B. (2000). Audio-visual and multimodal speech systems. In D. Gibbon & I. Mertins & R. K. Moore (Eds.), *Handbook of standards and resources for spoken language systems* (pp. 1-96). Dordrecht: Kluwer Academic Publishers.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*(6), 592-597.

Boersma, P., & Weenink, D. (2007). Praat: Doing phonetics by computer (Version 4.6.22). Amsterdam: University of Amsterdam.

Bouwman, A. G. G., Sturm, J., & Boves, L. (1999). *Incorporating confidence measures in the Dutch train timetable information system developed in the Arise Project.* Paper presented at the 24th International Conference on Acoustics, Speech, and Signal processing (ICASSP '99), Phoenix, USA.

Brebner, J., & Welford, A. (1980). Introduction: an historical background sketch. In A. Welford (Ed.), *Reaction times* (pp. 1–23). New York: Academic Press.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language, 34*(3), 383-398.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*(12), 1110-1123.

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences, 2*(7), 247-253.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics, 22*(2), 249-254.

Carlson, R., Hirschberg, J. B., & Swerts, M. G. J. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication, 46*(3-4), 326-333.

Carlson, R., & Swerts, M. G. J. (2003). *Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials.* Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS '03), Barcelona, Spain.

Carpenter, P., Jin, C., Wilson, D., Zhang, R., Bohus, D., & Rudnicky, A. I. (2001). *Is this conversation on track?* Paper presented at the 7th European Conference on Speech Communication and Technology (Eurospeech '01), Aalborg, Denmark.

Carroll, J. M., & Russell, J. A. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology, 70*(2), 205-218.

Caspers, J. (1998). Who's next? The melodic marking of question versus continuation in Dutch. *Language and Speech, 41*(3-4), 375-398.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). *Non-verbal cues for discourse structure.* Paper presented at the 39th Annual Meeting of the Association of Computational Linguistics (ACL '01), Toulouse, France.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). *About the relationship between eyebrow movements and F0 variations.* Paper presented at the 4th International Conference on Spoken Language Processing (ICSLP '96), Philadelphia, USA.

Cieri, C., Miller, D., & Walker, K. (2002). *Research methodologies, observations and outcomes in (conversational) speech data collection.* Paper presented at the 2nd International Conference on Human Language Technology Research (HLT '02), San Diego, USA

Clark, H. H. (1996). *Using language*. Cambridge (GB): Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259-294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39.

Cohn, J. F., & Katz, G. S. (1998). *Bimodal expression of emotion by face and voice.* Paper presented at the Workshop on Face/Gesture Recognition and Their Applications, the 6th ACM International Multimedia Conference (MM '98), Bristol, England.

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing, 2*(2), 1-12.

Cohn, J. F., Schmidt, K. L., Gross, R., & Ekman, P. (2002). *Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification.* Paper presented at the 4th International Conference on Multimodal User Interfaces (ICMI '02), Pittsburgh, USA.

Cohn, J. F., Xiao, J., Moriyama, T., Ambadar, Z., & Kanade, T. (2003). Automatic recognition of eye blinking in spontaneously occurring behavior. *Behavior Research Methods, Instruments, & Computers, 35*(3), 420-428.

Couper-Kuhlen, E. (1993). *English speech rhythm*. Philadelphia: Benjamins.

Danieli, M. (1996). *On the use of expectations for detecting and repairing human-machine miscommunication.* Paper presented at the Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication (AAAI '96), Portland, USA.

Dohen, M., Lœvenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). *Audiovisual Perception of Contrastive Focus in French* Paper presented at the International Conference on Audio-Visual Speech Processing (AVSP '03), St. Jorioz, France.

Doherty-Sneddon, G., Bonner, L., & Bruce, V. (2001). Cognitive demands of face monitoring: Evidence for visuospatial overload. *Memory & Cognition, 29*(7), 909-917.

Doughty, M. J. (2001). Consideration of three types of spontaneous eyeblink activity in normal humans: During reading and video display terminal use, in primary gaze, and while in conversation. *Optometry and Vision Science, 78*(10), 712-725.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. v. Cranach & K. Foppa & W. Lepenies & D. Ploog (Eds.), *Human ethology* (pp. 169-248). Cambridge (GB): Cambridge University Press.

Ekman, P. (1982). Methods for measuring facial action. In K. R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 45-90). New York: Cambridge University Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384-392.

Ekman, P. (2004). *Emotions Revealed: Recognizing faces and feelings to improve communication and emotional life*. New York: Owl books.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology, 58*(2), 342-353.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial expressions*. Englewood Cliffs, N.J.: Prentice Hall.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatis, I., Heider, K., Krause, R., LeCompte, W. A., Ricci-Bitti, P. E., Scherer, K. R., Tomita, M., & Tvaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*(4), 712-717.

Elfenbein, H. A., & Ambady, N. (2003). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology, 85*(2), 276-290.

Erickson, D., Fujimura, O., & Pardo, B. (1998). Articulatory correlates of prosodic control: Emotion versus emphasis. *Language and Speech, 41*(3-4), 399-417.

Fox, E., Lester, V., Russo, R., Bowles, R., Pichler, A., & Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition and Emotion, 14*(1), 61-92.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego, California: Academic Press.

Froeberg, S. (1907). The relation between the magnitude of stimulus and the time of reaction. *Archives of Psychology, 16*(8), 1-38.

Gagné, J.-P., Rochette, A.-J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37*(3-4), 213-230.

Gazzaniga, M. S., & Smylie, C. S. (1990). Hemispheric mechanisms controlling voluntary and spontaneous facial expressions. *Journal of Cognitive Neuroscience, 2*(3), 239-245.

de Gelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters, 260*(2), 133-136.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*(3), 289-311.

de Gelder, B., Vroomen, J., Pourtois, G., & Weiskrantz, L. (2000). Affective blindsight: Are we blindly led by emotions? *Trends in Cognitive Sciences, 4*(4), 126-127.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience, 25*(20), 5004-5012.

Goldberg, J., Ostendorf, M., & Kirschhoff, K. (2003). *The impact of response wording in error correction subdialogs*. Paper presented at the ISCA Workshop on Error Handling in Spoken Dialog Systems (EHSD '03), Château-d'Oex, Switzerland.

Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological Inquiry, 50*(3-4), 272-302.

Graf, H. P., Cosatto, E., Ström, V., & Huang, F. J. (2002). *Visual prosody: Facial movements accompanying speech.* Paper presented at the 5th International Conference on Automatic Face and Gesture Recognition (FGR '02), Washington, USA.

Granström, B., House, D., & Swerts, M. G. J. (2002). *Multimodal feedback cues in human-machine interactions.* Paper presented at the 1st Speech Prosody Conference (SP '02), Aix-en-Provence, France.

Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics, 21*(3), 501-529.

Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11*(6), 597-604.

Grossmann, T., Striano, T., & Friederici, A. (2006). Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science, 9*(3), 309-315.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*(4), 208-216.

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University press.

Hauser, M. D., & Akre, K. (2001). Asymmetries in the timing of facial and vocal expressions by rhesus monkeys: Implications for hemispheric specialization. *Animal Behaviour, 61*(2), 391-400.

Hearst, E. (1991). Psychology and nothing. *American Psychologist, 79*(5), 432-443.

van Heuven, V. J. (1994). Introducing prosodic phonetics. In C. Odé & V. J. van Heuven (Eds.), *Experimental Studies of Indonesian Prosody* (p. 214). Leiden University: Department of Languages and Cultures of South-East Asia and Oceania.

Hietanen, J. K., Manninen, P., Sams, M., & Rusakka, V. (2001). Does audiovisual speech perception use information about facial configuration? *European Journal of Cognitive Psychology, 13*(3), 395-407.

Hirschberg, J. B., Litman, D., & Swerts, M. G. J. (2001). *Identifying user corrections automatically in spoken dialogue systems.* Paper presented at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL '01), Pittsburgh, USA.

Hirschberg, J. B., Litman, D., & Swerts, M. G. J. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication, 43*(1-2), 155-175.

Hirschberg, J. B. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication, 36*(1-2), 31-43.

Hirst, D. (2004). *The phonology and phonetics of speech prosody: Between acoustics and interpretation.* Paper presented at the 2nd Speech Prosody Conference (SP '04), Nara, Japan.

Horstmann, G. (2002). Facial expressions of emotion: Does the prototype represent central tendency, frequency of instantiation, or an ideal? *Emotion, 2*(3), 297-305.

Humphreys, G. W., Donnelly, N., & Riddoch, M. J. (1993). Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia, 31*(2), 173-181.

Jakobs, E. B. (1998). *Faces and feelings in social context.* Unpublished Dissertation, Amsterdam University, Amsterdam.

Jordan, T. R., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43*(1), 107-124.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.

Kendon, A. (2001). Gesture as communication strategy. *Semiotica, 135*(1-4), 191-210.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition, 93*(1), 39-47.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? *Journal of Memory and Language, 48*(1), 16–32.

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature, 387*(6635), 767 - 768.

Koiso, H., Horiucho, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogues. *Language and Speech, 41*(3-4), 295-321.

Krahmer, E. J., & Swerts, M. G. J. (2004). More about brows: A cross-linguistic study via analysis-by-synthesis. In C. Pelachaud & Z. Ruttkay (Eds.), *From brows to trust: Evaluating embodied conversational agents* (pp. 191-216). Dordrecht: Kluwer Publishers.

Krahmer, E. J., Swerts, M. G. J., Theune, M., & Wegels, M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication, 36*(1-2), 133-145.

Krahmer, E. J., van Dorst, J., & Ummelen, N. (2004). Mood, persuasion and information presentation: The influence of mood on the effectiveness of persuasive digital documents. *Information Design Journal + Document Design, 12*(3), 40-52.

Kraut, R., & Johnson, R. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology, 37*(9), 1539–1553.

Lendvai, P., van den Bosch, A., Krahmer, E. J., & Swerts, M. G. J. (2002). *Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting.* Paper presented at the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics (ESSLLI '02), Trento, Italy.

Leppänen, J., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research, 69*(1-2), 22-29.

Leroy, L. (1984). The psychological reality of fundamental frequency declination. *Antwerp Papers in Linguistics, 40*, 1-102.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levow, G.-A. (2002). Adaptations in spoken corrections: Implications for models of conversational speech. *Speech Communication, 36*(1-2), 147-163.

Litman, D. J., Hirschberg, J. B., & Swerts, M. G. J. (2000). *Predicting automatic speech recognition performance using prosodic cues.* Paper presented at the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL '00), Seattle, Washington.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mackie, D., & Worth, L. (1989). Processing deficits and the mediation of positive affect in persuasion. *Journal of Personality and Social Psychology, 57*(1), 27-40.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion, 14*(3), 195-214.

Matsumoto, D., & Ekman, P. (1989). American-Japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion, 13*(2), 143-157.

Maynard, S. K. (1987). Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics, 11*(5), 589-606.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics, 32*(7), 855-878.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: Chicago University Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: Chicago University Press.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science, 15*(2), 133-137.

Nakano, M., & Hazen, T. J. (2003). *Using untranscribed user utterances for improving language models based on confidence scoring.* Paper presented at the 8th European Conference on Speech Communication and Technology (Eurospeech '03), Geneva, Switzerland.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003, July 7-12). *Towards a model of face-to-face grounding.* Paper presented at the 41st Annual Meeting on Association for Computational Linguistics (ACL '03), Sapporo, Japan.

Nooteboom, S. G. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 640-673). Blackwell, Oxford.

Novick, D. G., Hansen, B., & Ward, K. (1996). *Coordinating turn-taking with gaze.* Paper presented at the 4th International Conference on Spoken Language Processing (ICSLP '96), Philadelphia, USA.

Oviatt, S., MacEachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*(2), 1-23.

Petajan, E. D. (1985). *Automatic lipreading to enhance speech recognition.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '85), Atlanta, Georgia.

Picard, R. W., & Klein, J. (2002). Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers, 14*(2), 141-169.

de Pijper, J. R., & Sanderman, A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America, 96*(4), 2037-2047.

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Cognitive Neuroscience, 11*(6), 1329-1333.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, S. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*(6), 2956-2970.

Rietveld, A. C. M., & van Heuven, V. J. (2001). *Algemene fonetiek*. Bussum: Coutinho.

Rinn, W. E. (1984). The neuropsychology of facial expressions: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin, 95*(1), 52-77.

Rinn, W. E. (1991). Neuropsychology of facial expression. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 3-30). Cambridge: Cambridge University Press.

Ross, E. D. (2000). Affective prosody and the aprosodias. In M.-M. Mesulam (Ed.), *Principles of behavioral and cognitive neurology* (pp. 316-331). New York: Oxford University Press.

de Ruiter, J. P., Miterrer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515-535.

Russell, J. A., Bachorowski, J., & Fernández-Dols, J. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*, 329–349.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language, 50*(4), 696-735.

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

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology, 44*, 3-24.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*(2), 69-78.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47*(3), 277-287.

Sengers, P. (1999). *Designing comprehensible agents.* Paper presented at the 16th International Joint Conference of Artificial Intelligence (IJCAI '99), Stockholm, Sweden.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193-247.

Silverman, K., & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In J. Kingston & M. Beckman (Eds.), *Laboratory Phonology, Vol I: Between the grammar and physics of speech* (pp. 71-106). Cambridge: Cambridge University Press.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*(1), 25-38.

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience, 2*(2), 191-196.

Swerts, M. G. J. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America, 101*(1), 514-521.

Swerts, M. G. J. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

Swerts, M. G. J., Bouwhuis, D. G., & Collier, R. (1994). Melodic cues to the perceived finality of utterances. *Journal of the Acoustical Society of America, 96*(4), 2064-2075.

Swerts, M. G. J., Collier, R., & Terken, J. (1994). Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication, 15*(1-2), 79-90.

Swerts, M. G. J., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech, 37*(1), 21-43.

Swerts, M. G. J., Krahmer, E., Barkhuysen, P. N., & van de Laar, L. (2003). *Audiovisual cues to uncertainty.* Paper presented at the ISCA Workshop on Error Handling in Spoken Dialog Systems (EHSD '03), Chateau-d'Oex, Switzerland.

Teichner, W., & Krebs, M. (1974). Laws of visual choice reaction time. *Psychological Review, 81*(1), 75-98.

Ulrich, R., Rinkenauer, G., & Miller, J. (1998). Effects of stimulus duration and intensity on simple reaction time and response force. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 915-928.

Valstar, M., Pantic, M., Ambadar, Z., & Cohn, J. (2006). *Spontaneous vs. posed facial behavior: Automatic analysis of brow actions.* Paper presented at the 8th International Conference on Multimodal Interfaces (ICMI '06), Banff, Canada.

Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research & Therapy, 6*(4), 473-482.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2000). *Why conversational agents should catch the eye.* Paper presented at the ACM Conference on Human Factors in Computing Systems (CHI '00), Den Haag, The Netherlands.

Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia, 45*(1), 174-194.

Wade, E., Shriberg, E., & Price, P. (1992). *User behaviors affecting speech recognition.* Paper presented at the 2nd International Conference on Spoken Language Processing (ICSLP '92), Banff, Canada.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3-28.

Wagner, H. L. (1997). Methods for the study of facial behavior. In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 31-54). Cambridge: Cambridge University Press.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1998). Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language, 12*(3), 317-347.

Walley, A. C., Michela, V. L., & Wood, D. R. (1995). The gating paradigm: Effects of presentation format on spoken word recognition by children and adults. *Perception & Psychophysics, 57*(3), 343–351.

Wang, S. B., Demirdjian, D., & Darrell, T. (2007). *Detecting communication errors from visual cues during the system's conversational turn.* Paper presented at the 9th International Conference on Multimodal Interfaces (ICMI '07), Nagoya, Japan.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32*(8), 1177–1207.

Weaver, W., & Shannon, C. E. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Wells, G. (1913). The influence of stimulus duration on RT. *Psychological Monographs, 15*, 1-69.

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*(4), 557-580.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707-1717.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics, 38*(10), 1559–1579.

Wilting, J., Krahmer, E. J., & Swerts, M. G. J. (2006). *Real vs. acted emotional speech.* Paper presented at the 9th International Conference on Spoken Language Processing (ICSLP '06), Pittsburgh, USA.

Summary

This thesis addresses the role of the modality in conveying prosody in audiovisual speech. Although a lot is known about how prosody is expressed in the voice, less is known about how prosody is expressed in the face. There are reasons to believe that the combination of displaying prosody in the face as well as in the voice, i.e. a bimodal, or audiovisual, expression of prosody, may be more effective for speech perception than when listeners have access to a single modality.

Three studies are discussed, each covering a different component of setting a common ground (i.e. the cooperation between speaker and listener to ensure mutual understanding (Clark, 1996, p.12; Stalnaker, 1978, in Clark & Schaefer, 1989)): (1) how dialogue participants provide feedback, e.g. how they signal and detect communication problems, (2) how participants regulate turn-taking, e.g. how they display and detect end-of-utterance marking, and (3) how participants display and perceive emotions. In each of these three subdomains we focused on the role of the visual and the auditory modality, the relative importance of each modality, and possible interactions between them. In all three studies, we presented the recorded stimulus material in different modalities. These modalities were a bimodal, audiovisual condition (i.e. the original recording), or a unimodal condition, which was either vision-only (only the visual signal, i.e. facial expressions, stored) or audio-only (only the auditory signal, i.e. the voice, stored).

Chapter 2 described research into audiovisual cues to communication problems in interactions between users and a spoken dialogue system. The study consisted of two parts. First, we described a series of three perception experiments in which participants were offered film fragments (without any dialogue context) of speakers interacting with a spoken dialogue system, offering a train travel advice. In half of these fragments, the speaker is or becomes aware of a communication problem. Participants had to determine by forced choice which are the problematic fragments. The perception tests reflected different phases in the dialogue: verification questions of the system (e.g. "*So you want to travel to Amsterdam?*"), negations ("*no*") and slot fillers (e.g. "To which station do you want to travel?" - "*Utrecht*") on the part of the speaker. In all three tests, participants were capable of performing this task to some extent, but with varying levels of correct classifications. The

negations were most difficult, the verification questions were classified best. Because the verification questions showed people listening to a system's question, it could be stated that participants perform better in the 'vision-only' condition (verification questions) than in the bimodal conditions (negations and slot-fillers).

Second, we reported results of an observational analysis in which we first attempted to relate the perceptual results to features of the stimuli presented to participants, and second to find out which visual features actually are potential cues for error detection. Because the negations were the shortest stimuli, we speculated that in longer stimuli there may be more cues available. Also, the classifications differed across speakers, which led us to the question whether it is possible that different speakers show different cues. On the basis of these two questions we decided to do an additional observational analysis, in which we showed that more problematic contexts lead to more dynamic facial expressions, in line with earlier claims that communication errors lead to marked speaker behavior. Both hyperarticulation (i.e. exaggerated speech, e.g. "AM....ste:rr..dam") and the amount of visual variation (which is the sum of several individual facial features) played a role. Also, the presence of these features influenced the perception of problems. This chapter thus showed that visual information from a user's face is potentially beneficial for problem detection, and that a system may use this information by automatic facial tracking, in order to monitor the user's frustration and the concurrent presence of errors.

Chapter 3 discussed how audiovisual prosody is used to signal the end of an utterance, which may play a role in the fluency of turn-taking. Speaker utterances were collected via a novel semi-controlled production experiment, in which participants provide lists of words in an interview setting (e.g. "What are the colors of the Dutch flag?" - "*Red.. white.. blue*", or "What are the odd numbers between ten and zero, in descending order?" - "*Nine.. seven.. five.. three.. one*"). These data were used in two perception experiments, which systematically compared responses to unimodal (audio-only and vision-only) and bimodal (audiovisual) stimuli. The first experiment was a reaction time experiment, in which participants had to indicate when they think the end of the utterance is reached. This experiment revealed that participants are significantly quicker in end-of-utterance detection when confronted with bimodal or audio-only stimuli, than for vision-only stimuli. Also, short stimuli (e.g. "*seven*") led to longer reaction times than long stimuli (e.g. "*nine.. seven*"). No significant differences in reaction times were found between the bimodal and audio-only condition, and therefore a second experiment was conducted. The second experiment was a classification experiment, in which participants had to indicate whether a fragment is final or not. This experiment showed that participants perform significantly better in the bimodal condition than in the two unimodal ones. Also, short stimuli were more difficult to classify than long stimuli. Further, non-final stimuli (e.g. "*nine.. seven*" versus "three.. one") were classified better, but there was a small response bias. Both the first and the second experiment revealed interesting differences between speakers in the various conditions, which indicates that some speakers are more expressive in the visual and others in the auditory modality.

In an additional observational analysis, the results of these perception tests were linked to features in the stimuli, in order to find out which features are potential cues for end-of-utterance detection. We showed that several auditory as well as visual cues seemed to play a role in the judgment of finality. Further, there were large differences between speakers in the amount and type of features displayed, also depending on the place in the utterance. It was suggested that many of these features may be rhythmically distributed over the different phases within an utterance (e.g. a rhythmic pattern of nodding on words, or diverting and returning the head). This chapter thus showed that speakers also employ visual cues, apart from auditory cues such as intonation, rhythm, and pausing, to indicate that they approach the end of their utterance.

In Chapter 4 we investigated how audiovisual emotional speech is displayed. We collected video clips of emotional utterances via a variant of the well-known Velten method. More specifically, we recorded speakers who displayed positive or negative emotions, which were congruent or incongruent with the (emotional) lexical content of the uttered sentence (e.g. "God, I feel great!" uttered in a 'happy' or 'unhappy' way). In order to test this, we conducted two experiments. The first experiment was a perception experiment in which Czech participants, who did not speak Dutch, rated the perceived emotional state of Dutch speakers in a bimodal (audiovisual) or a unimodal (audio- or vision-only) condition on a scale from 1 to 7. This experiment showed no overall differences between the modalities. It was found that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, where the difference between congruent and incongruent emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between congruent and incongruent emotions were found for the audio-only condition, which suggests that 'posing' an incongruent emotion has a particularly strong effect on the spoken realization of emotions. The second experiment used a gating paradigm to test the recognition speed for various emotional expressions from a speaker's face. In this experiment participants were presented with the same clips as experiment 1, but this time presented vision-only. The clips were shown in successive segments (gates) which increase in length (e.g. from "Go.." to "God I feel gr.."). Results showed that participants are surprisingly accurate in their recognition of the various emotions, as they already reach high recognition scores in the first gate (after only 160 milliseconds). Interestingly, the recognition scores rose faster for positive than for negative conditions. Finally, the gating results suggested that incongruent emotions were perceived as more intense than congruent emotions, as the former get more extreme recognition scores than the latter, already after a short period of exposure.

In an additional observational analysis, the results of these perception tests were linked to features in the stimuli, in order to find out which features are potential cues for emotion perception. The observational analysis showed that the occurrence of three visual features differed depending on the valence of the emotion (i.e. positive or negative) and whether the speaker is experiencing an emotion congruent with the lexical content. It was also shown that the occurrence of these features was related to the perceived emotional state. This chapter thus showed that people classify emotions in a different way across different modalities, depending on the valence of the emotion and whether the speaker is experiencing an emotion congruent with the lexical content.

Chapter 5 presented the main results of the three studies. These findings were discussed in the light of the assumptions on the basis of which we designed the experiments. We started with the assumption that prosody is multimodal in its nature, that the role of the sender as well as of the receiver is important, and that we should investigate natural interactions. The first conclusion was that all three chapters suggested that it is possible to interpret signals from the visual modality only, which makes it likely that a phenomenon as 'visual' prosody exists. The results also suggested that integration occurs to some extent (e.g. in detecting end-of-utterance). The auditory and the visual modality are equally informative, but they can complement each other: when information is missing in one modality the participants can turn towards the other modality. Further, the information value of a modality depends on other factors, such as the speaker involved, characteristics of the stimuli and the role that the utterance has within the dialogue.

The second conclusion was that speakers display a wide spectrum of behaviors. They display auditory and visual cues, large movements such as posture changes as well as subtle movements such as eye movements. The distribution of cues over the auditory versus visual modality seemed to be speaker-dependent. We did not further explore these speaker differences, but suggested that they may have been caused by factors as sex, age, personality and culture.

The third conclusion was that the evoked expressions were natural, but it was not completely clear to what extent we can generalize the results to daily life situations, due to the 'observer's paradox'.

We also made some suggestions for future research, concentrating upon extending the context (social and cross-cultural differences), the human factor (role of the receivers) and the level of detail (refinement of cue measurement and timing).

Samenvatting

Dit proefschrift behandelt de rol van de modaliteit in het uitdrukken van prosodie[25] in audiovisuele spraak. Hoewel er veel bekend is over hoe prosodie uitgedrukt wordt in de stem, is er minder bekend over hoe prosodie uitgedrukt wordt in het gezicht. Er zijn redenen om aan te nemen dat de combinatie van het uitdrukken van prosodie in zowel het gezicht als in de stem, d.w.z. een bimodale, of audiovisuele, expressie van prosodie, effectiever kan zijn voor de perceptie van spraak dan wanneer luisteraars toegang hebben tot een enkele modaliteit.

Er worden drie studies beschreven, die elk een verschillend onderdeel beschrijven in het vestigen van een gemeenschappelijke basis ('*common ground*', d.w.z. de samenwerking tussen spreker en luisteraar om zich te verzekeren van een wederzijds begrip (Clark, 1996, p.12; Stalnaker, 1978, in Clark & Schaefer, 1989)): (1) hoe deelnemers feedback geven in een dialoog, bijv. hoe ze communicatieproblemen uitdrukken en opmerken, (2) hoe deelnemers de beurtwisseling reguleren, bijv. hoe ze het einde van een uiting uitdrukken en opmerken, en (3) hoe deelnemers emoties uitdrukken en opmerken. In elk van deze drie subdomeinen richtten we ons op de rol van de visuele en de auditieve modaliteit, de relatieve belangrijkheid van elke modaliteit, en de mogelijke interacties ertussen. In alle drie de studies presenteerden we het stimulusmateriaal in verschillende modaliteiten. Deze modaliteiten bestonden uit een bimodale, audiovisuele conditie (d.w.z. de originele opname), of uit een unimodale conditie, die ofwel alleen visueel was (alleen het visuele signaal, d.w.z. gezichtsuitdrukkingen, waren opgeslagen) ofwel alleen auditief (alleen het auditieve signaal, d.w.z. de stem, was opgeslagen).

Hoofdstuk 2 beschrijft onderzoek naar audiovisuele kenmerken van communicatieproblemen in interacties tussen gebruikers en een gesproken dialoogsysteem. De studie bestond uit twee delen. Ten eerste beschreven we een serie van drie perceptie-experimenten waarin deelnemers werden blootgesteld aan filmfragmenten (zonder de context van de dialoog) van sprekers die met een gesproken dialoogsysteem

---

[25] Prosodie is de combinatie van oa. de melodie in de stem (intonatie), en de duur en het ritme van klanken in de spraak.

communiceerden, welke een reisadvies voor de trein uitbracht. In de helft van deze fragmenten was of werd de spreker zich bewust van een communicatieprobleem. Deelnemers moesten middels een gedwongen keuze aangeven welke de problematische fragmenten waren. De perceptietesten weerspiegelden verschillende fases in een dialoog: verificatievragen van het systeem (bijv. '*Dus u wilt naar Amsterdam reizen?*'), ontkenningen ("*nee*") en plaatsnamen (bijv. "Naar welk station wilt u reizen?" - "*Utrecht*") aan de kant van de spreker. In alle drie de testen waren deelnemers in staat dit in zekere mate te doen, maar met verschillende niveaus van correcte classificaties. De ontkenningen waren het moeilijkst, de verificatievragen werden het best geclassificeerd. Omdat de verificatievragen mensen toonden die naar een vraag van het systeem aan het luisteren waren, kunnen we stellen dat deelnemers beter presteerden in een 'alleen visuele' conditie (verificatievragen) dan in de bimodale condities (ontkenningen en plaatsnamen).

Ten tweede rapporteerden we de resultaten van een observatieanalyse waarin we eerst probeerden de perceptuele resultaten aan kenmerken in de stimuli die aan de deelnemers gepresenteerd waren te relateren, en vervolgens uit probeerden te vinden welke visuele kenmerken daadwerkelijk mogelijke cues voor foutdetectie waren. Omdat de ontkenningen ook de kortste stimuli waren, speculeerden we dat er in langere stimuli meer kenmerken beschikbaar zouden kunnen zijn. Tevens verschilden de classificaties over sprekers, wat ons op de vraag bracht of het mogelijk is dat verschillende sprekers andere kenmerken vertonen. Op de basis van deze twee vragen besloten we een additionele observatieanalyse te doen, waarin we aantoonden dat contexten die problematischer zijn leidden tot dynamischere gezichtsuitdrukkingen, in overeenstemming met eerdere claims dat communicatie fouten leiden tot gemarkeerd sprekergedrag. Zowel hyperarticulatie (d.w.z. overdreven spraak, bijv. "AM....ste:rr..dam") als de mate van visuele variatie (wat de som is van verschillende individuele gezichtsuitdrukkingen) speelden een rol. Ook beïnvloedde de aanwezigheid van deze kenmerken de perceptie van problemen. Dit hoofdstuk toonde aldus aan dat visuele informatie op het gezicht van een gebruiker mogelijk nuttig kan zijn voor foutdetectie, en dat een systeem deze informatie zou kunnen gebruiken door het automatisch volgen van het gezicht, teneinde de frustratie van de gebruiker en de daarmee samengaande aanwezigheid van fouten te controleren.

Hoofdstuk 3 besprak hoe audiovisuele prosodie wordt gebruikt om het verloop van een uiting te vertonen, wat een rol kan spelen in de vloeiendheid van de beurtwisseling. Er werden sprekersuitingen verzameld via een nieuw semi-gecontroleerd productie-experiment, waarin deelnemers lijsten van woorden verschaffen in een interview setting (bijv. "Wat zijn de kleuren van de Nederlandse vlag?" - "*Rood.. wit.. blauw*", of "Wat zijn de oneven getallen tussen tien en nul, in afnemende volgorde?" - "*Negen.. zeven.. vijf.. drie.. één*"). Deze gegevens werden gebruikt in twee perceptie-experimenten, die systematisch reacties op unimodale (alleen auditieve en alleen visuele) en bimodale (audiovisuele) stimuli vergeleken. Het eerste experiment was een reactietijdexperiment, waarin deelnemers moesten aangeven wanneer ze denken dat het einde van een uiting is bereikt. Dit experiment onthulde dat deelnemers significant sneller zijn in het detecteren van het einde van een uiting wanneer ze geconfronteerd worden met bimodale of alleen auditieve stimuli, dan met alleen visuele stimuli. Tevens leidden korte stimuli (bijv. "*zeven*") tot langere reactietijden dan langere stimuli ("*negen.. zeven*"). Er werden geen significante verschillen in reactietijden gevonden tussen de bimodale en alleen auditieve conditie, en daarom werd er een tweede experiment uitgevoerd. Het tweede experiment was een classificatie-experiment, waarin deelnemers moesten aangeven of een fragment finaal was of niet. Dit experiment toonde aan dat deelnemers significant beter presteren in de bimodale conditie dan in de twee unimodale condities. Tevens waren korte stimuli moeilijker te classificeren dan lange stimuli. Verder werden niet-finale stimuli (bijv. "*negen.. zeven*" versus "drie.. één") beter geclassificeerd, maar er was een kleine response bias. Zowel het eerste als het tweede experiment onthulden interessante verschillen tussen sprekers in de verschillende condities, wat aangeeft dat sommige sprekers expressiever zijn in de visuele en andere in de auditieve modaliteit.

In een additionele observatieanalyse werden de resultaten van de perceptietesten gelieerd aan kenmerken in de stimuli, om uit te vinden welke kenmerken mogelijke cues zijn voor het detecteren van het einde van een uiting. We toonden aan dat verscheidene auditieve en visuele kenmerken een rol leken te spelen in het beoordelen van finaliteit. Verder waren er grote verschillen tussen sprekers in de hoeveelheid en type van vertoonde kenmerken, tevens afhankelijk van de plaats in de uiting. Er werd gesuggereerd dat veel van deze kenmerken ritmisch verdeeld zouden kunnen zijn over de verschillende fases in een uiting (bijv. een ritmisch patroon van knikken op woorden, of het wegdraaien en het terugdraaien van het hoofd). Dit hoofdstuk toonde aldus aan dat sprekers ook visuele kenmerken gebruiken, apart van auditieve kenmerken zoals intonatie, ritme en pauzeren, om aan te geven dat ze het einde van hun uiting naderen.

In hoofdstuk 4 beschrijven we hoe audiovisuele prosodie wordt vertoond in emotionele spraak. We verzamelden videoclips van emotionele uitingen via een variant van de bekende 'Velten methode'. In concreto, we verzamelden sprekers die positieve of negatieve emoties vertoonden, die in overeenstemming (congruent) of in conflict (incongruent) waren met de (emotionele) lexicale inhoud van de geuite zin (bijv. "God, ik voel me geweldig!" geuit op een 'blijde' of een 'niet blijde' manier). Om dit te testen voerden we twee experimenten uit. Het eerste experiment was een perceptie-experiment waarin Tsjechische deelnemers, die geen Nederlands spraken, de waargenomen emotionele toestand van Nederlandse sprekers in een bimodale (audiovisuele) of een unimodale (alleen auditieve of visuele) conditie moesten waarderen op een schaal van 1 tot 7. Dit experiment toonde geen significante verschillen tussen de modaliteiten aan. Er werd gevonden dat incongruente emotionele spraak leidt tot significant extremere scores van de waargenomen emotionele toestand dan congruente emotionele spraak, waarbij het verschil tussen congruente en incongruente emotionele spraak groter is voor de negatieve dan voor de positieve condities. Het is interessant op te merken dat de grootste algemene verschillen tussen congruente en incongruente emoties werden gevonden in de alleen auditieve conditie, wat suggereert dat het 'acteren' van een incongruente emotie een bijzonder sterk effect heeft op de gesproken realisatie van emoties. Het tweede experiment gebruikt een gating paradigma (d.w.z. het gebruik van 'vensters' danwel 'hekken' of 'horden') om de herkenningssnelheid te testen voor verscheidene emotionele expressies van het gezicht van een spreker. In dit experiment werden deelnemers blootgesteld aan dezelfde fragmenten als in experiment 1, maar deze keer alleen visueel gepresenteerd. De fragmenten werden vertoond in opeenvolgende segmenten (gates) die in duur toenamen (bijv. van "Go.." tot "God, ik voel me gew.."). De resultaten toonden aan dat deelnemers verrassend goed zijn in hun herkenning van de verschillende emoties, aangezien ze al hoge herkenningsscores bereiken in het eerste segment (na slechts 160 milliseconden). Het is interessant te vermelden dat de herkenningscores sneller stegen voor positieve dan voor negatieve condities. Als laatste suggereerden de resultaten van dit experiment dat incongruente emoties als intenser worden waargenomen dan congruente emoties, aangezien de eerstgenoemde extremere herkenningsscores krijgen dan de laatstgenoemde, al na een korte tijd van blootstelling.

In een additionele observatieanalyse werden de resultaten van deze perceptietesten gelieerd aan kenmerken in de stimuli, om uit te vinden welke kenmerken potentiële cues zijn voor de perceptie van emoties. De observatieanalyse toonde aan dat het vóórkomen van drie visuele kenmerken verschilde al naar gelang de waarde van de emotie (d.w.z. positief of negatief) en of de spreker een emotie ervaart die congruent is met de lexicale inhoud. Er werd ook aangetoond dat het vóórkomen van deze kenmerken gerelateerd was aan de waargenomen emotionele toestand. Dit hoofdstuk toonde aldus aan dat mensen emoties verschillend classificeren over verschillende modaliteiten, afhankelijk van de waarde van de emotie en van of de spreker een emotie ervaart die congruent is met de lexicale inhoud.

Hoofdstuk 5 presenteerde de voornaamste resultaten uit de drie studies. De bevindingen werden besproken in het licht van de assumpties op de basis waarvan we de experimenten ontwierpen. We begonnen met de assumptie dat prosodie multimodaal van aard is, dat de rol van zowel de zender als de ontvanger belangrijk is, en dat we natuurlijke interacties zouden moeten bestuderen. De eerste conclusie was dat alle drie de hoofdstukken suggereerden dat het mogelijk is dat om signalen uit alleen de visuele modaliteit te interpreteren, wat het waarschijnlijk maakt dat een fenomeen als 'visuele prosodie' bestaat. De resultaten suggereerden ook dat integratie in een bepaalde mate voorkomt (bijv. in het detecteren van het einde van een uiting). De auditieve en de visuele modaliteit zijn even informatief, maar ze kunnen elkaar aanvullen: waneer er informatie uit de ene modaliteit mist kan de deelnemer zich tot de andere modaliteit wenden. Verder is de informatiewaarde van een modaliteit afhankelijk van andere factoren, zoals de betrokken spreker, kenmerken van de stimuli, en de rol die de uiting heeft binnen de dialoog.

De tweede conclusie was dat sprekers een breed spectrum aan gedrag vertonen. Ze vertonen auditieve en visuele kenmerken, zowel grote bewegingen zoals houdingswisselingen als subtiele bewegingen zoals oogbewegingen. De verdeling van kenmerken over de auditieve versus de visuele modaliteit leek afhankelijk van de spreker te zijn. We onderzochten deze spreker verschillen niet verder, maar suggereerden dat ze veroorzaakt zouden kunnen zijn door factoren als sekse, leeftijd, persoonlijkheid en cultuur. De derde conclusie was dat de uitgelokte expressies natuurlijk waren, maar het was niet helemaal duidelijk in hoeverre we de resultaten kunnen generaliseren naar situaties in het dagelijkse leven, vanwege de 'observator's paradox'.

We deden ook enkele suggesties voor toekomstig onderzoek, waarbij we ons concentreerden op het uitbreiden van de context (sociale en culturele verschillen), de menselijke factor (de rol van de ontvangers), en het nivo van detail (verfijning van het meten van de cues en de timing).

Curriculum Vitae

Pashiera Barkhuysen was born on March 6th 1972 in Dwingeloo, The Netherlands. She studied Cognitive Psychology at the Radboud University in Nijmegen and specialized in Psycholinguistics. Her master thesis was about the influence of limitations in the working memory on the production of subject-verb agreement. Subsequently, Pashiera worked for 5 years in the IT business as a software developer and web developer. In 2003, Pashiera started her PhD research at the Communication and Cognition Group at Tilburg University. She participated in the project "Functions of Audiovisual Prosody" which was a part of a NWO research programme.

Scientific publications

The following is a list of publications that have been published or submitted by the author, ordered according to publication date. For all papers where I am the first author, I have been the main researcher and writer.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (accepted). Cross-modal and incremental perception of audiovisual emotional speech. *Language and Speech*.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2008). The interplay between auditory and visual cues for end-of-utterance detection. *The Journal of the Acoustical Society of America*, *123*(1), pp. 354-365.

Hartsuiker, R. J., & Barkhuysen, P. N. (2006). Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes*, *21*(3), pp. 181-204.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2005), Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication*, *45*(3), pp. 343-359.

Proceedings

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2007). *Incremental perception of acted and real emotional speech*, Paper presented at the 10th European Conference on Speech Comunication and Technology (ICSLP '07), Antwerp, Belgium.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2007). *Cross-modal perception of emotional speech*, Paper presented at the 16th International Congress of Phonetic Sciences (ICPhS '07), Saarbrücken, Germany.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2006). *How auditory and visual prosody is used in end-of-utterance detection*, Paper presented at the 9th International Conference on Spoken Language Processing (ICSLP '06), Pittsburgh, USA.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2005). *Audiovisual cues to finality*, Paper presented at the 5th International Conference on Methods and Techniques in Behavioral Research (MB '05), Wageningen, Netherlands.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2005). *Predicting end of utterance in multimodal and unimodal conditions*, Paper presented at the 9th European Conference on Speech Communication and Technology (ICSLP '05), Lisbon, Portugal.

Barkhuysen, P. N., Krahmer, E. J. & Swerts, M. G. J. (2004). *Audiovisual perception of communication problems*, Paper presented at the 2nd Speech Prosody Conference (SP '04), Nara, Japan.

Swerts, M. G. J., Krahmer, E. J., Barkhuysen, P. N., & van de Laar, L. (2003). *Audiovisual cues to uncertainty*, Paper presented at the ISCA Workshop on Error Handling in Spoken Dialog Systems (EHSD '03), Chateau-d'Oex, Switzerland.

Hartsuiker, R. J., Barkhuysen, P. N., & Watson, M. (2003). *Grammatical encoding as a resource-limited process?* Poster presented at the 16th Annual CUNY Conference on Human Sentence Processing (CUNY '03), Boston, USA.

Hartsuiker, R. J., & Barkhuysen, P. N. (2001). *The production of subject-verb agreement and verbal working memory*. Paper presented at the 7th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP '01), Saarbrücken, Germany.