# SEMANTIC FRAMING OF SPEECH

## Emotional and Topical Cues in Perception of Poorly Specified Speech

Björn Lidestam

FACULTY OF ARTS AND SCIENCES
LINKÖPING UNIVERSITY

Linköping Studies in Education and Psychology No. 94
Linköping University, Department of Behavioural Sciences

Linköping 2003

LINKÖPING UNIVERSITY
Department of Behavioural Sciences
S-581 83 Linköping
Sweden

*Semantic Framing of Speech:*
*Emotional and Topical Cues in Perception of Poorly Specified Speech*

Printed by UniTryck, Linköping, 2003
ISRN LIU-IBV-STU--94--SE
ISBN 91-7373-754-2
ISSN 1102-7517

# PREFACE

This thesis is in the middle of many fields of research that all may contribute with valuable perspectives on the present phenomena under study. Cognitive psychology is my ”home field”, but I have also used theories and methods from, foremost, psychology of perception, psycholinguistics, psychology of emotion, and social psychology. I will also discuss empirical findings from these fields. In doing this, I am well aware of that some common terms have somewhat different meanings and connotations, and that they also are used differently depending on the field. I have therefore tried to the best of my knowledge to define the terms as they are used in the thesis.

The following five studies, which will be referred to in the text by Roman numerals, constitute the basis for the thesis.

I. Lyxell, B., Johansson, K., Lidestam, B., & Rönnberg, J. (1996). Facial expressions and speechreading performance. *Scandinavian Audiology*, *25*, 97–102.

II. Lidestam, B., Lyxell, B., & Andersson, G. (1999). Speech-reading: Cognitive predictors and displayed emotion. *Scandinavian Audiology*, *28*, 211–217.

III. Lidestam, B., Lyxell, B., & Lundeberg, M. (2001). Speech-reading of synthetic and natural faces: Contextual cueing and mode of presentation. *Scandinavian Audiology*, *30*, 89–94.

IV. Lidestam, B. (2002). Effects of displayed emotion on attitude and impression formation in visual speech-reading. *Scandinavian Journal of Psychology*, *43*, 261–268.

V. Lidestam, B., & Beskow, J. (2003). *Effects of emotional and scriptural cues in auditory and visual speech perception*. Manuscript submitted for publication.

# ACKNOWLEDGEMENTS

My work has not been done by just pressing keys on my computer keyboard; other people have of course been involved in various ways. Björn Lyxell, my supervisor, was solid as rock all along the way. He encouraged me, provided invaluable insights, and, among other things, coined a phrase after me. This work could definitely not have been done without him. Jan Andersson scrutinised a previous version of the manuscript thoroughly, and provided helpful advice on both scientific and structural aspects. Ulrik Johansson and Ulla-Britt Persson checked the language on many occasions. Jerker Rönnberg was a source of inspiration with his enthusiastic leadership of our CCDD division and his cycling to Finspång. He was also very good to have for discussing ideas with and for picking up various forms of information from whenever needed. My colleagues at the division also helped with the thesis in many ways, and contributed to my general well-being by providing a good atmosphere. Ulf Andersson, Stefan Gustafson, Ulrich Olofsson, Charlotte Alm, Henrik Danielsson, Mary Rudner, Janna Ferreira, Anna Levén, and Stefan Samuelsson all contributed in some way to the work. Jonas Beskow, Gerhard Andersson, Anneli Sepa, Johan Söderquist, Lauritz Brännström, Karina Johansson, Magnus Lundeberg, Thomas Lunner, Stig Arlinger, Bertil Lyberg, and Mathias Henningsson are colleagues from other divisions, departments and universities and who also contributed. Last, but not least, all kind participants in the experiments deserve special recognition, as their contribution was an absolute prerequisite for the thesis. I hope not to have deterred any of you from participating in psychological experiments in the future.

The research and writing was financed by a grant from the Swedish Agency for Innovation Systems, formerly the Swedish Transport & Communications Research Board (1997-0603), awarded to Björn Lyxell. Gunilla Lundén helped us in the contact with the board and the agency whenever needed. Stiftelsen Tysta Skolan allowed scholarships for travelling to two conferences on auditory–visual speech processing.

This thesis is dedicated to Helene, Adina, and Gunnar. It could not have been done without Helene; neither could Adina, nor Gunnar. *I owe you*.

Linköping, September, 2003

Björn Lidestam

# CONTENTS

# INFORMATION PROCESSING: THE DOMAIN OF COGNITIVE PSYCHOLOGY

## Sources of Information

When we sense, perceive and think, we process information. When you read this, the physical properties of the print vis-a-vis the physical properties of the paper are transmitted as light, sensed, and recognised as features. The combined features, in turn, are perceived as patterns: letters and words. The organisation of input information as patterns is fundamental to perception, as postulated already by the Gestaltists (e.g., Koffka, 1935). Other examples of fundamental principles of perception are, according to the Gestaltists, that a Gestalt (i.e., percept) is more than the sum of its parts, due to the laws of perceptual organisation; and that a Gestalt may be construed on the basis of figure and background (Koffka, 1935).

In order for feature matching and pattern recognition to be possible, the stimuli must be compared to information that you have stored. However, you do not match optical features and merely recognise the letters as patterns that you have seen before; you associate the letters with sounds, combine them into words and sentences that *mean* something. The point is, that reading this, effortless as it may be, involves a great deal of *information*: from the few milligrammes of printing ink before you, and from your store of information within your brain. All of this information also has to be *processed* in an orderly and efficient fashion, or else you will not have met the basic requirements for understanding the text up to this point. If you also have understood the meaning of this text, you must, for example, have accessed stored knowledge of the English language with its rules and vocabulary.

When this text enters your consciousness, it can be called a *stimulus*. Stimuli are pieces of information that come from outside the unit that processes the information, and that may elicit responses within the unit. Right now, you are the unit that processes the text, which is before you. In order for you to be able to respond adequately to the text (i.e., understand it), you need information from a *lexicon*[1]. The lexicon can be thought of as your store of information that allows you to successfully interact with the complex world around you, and it contains, among other things, *concepts*. Concepts refer to mental representations of ”how things are related or categorised” (Harley, 1995, p. 176). Now, when you read this, information must flow in two directions: from the stimuli, and from your lexicon. Your processing of the information can therefore be divided into *stimulus-driven* and *conceptually driven* processing, or *bottom-up* and *top-down* processing, respectively.

Sometimes top-down influence, such as from concepts, makes a small difference to what we perceive, and sometimes it makes a large difference. For example, you can probably read this sentence without difficulty, and had you not read the sentences preceding it, you would still have no difficulty recognising the words, provided that you have good vision and normal reading skills. On the other hand, if either the text from here on was in bad handwriting, or your sight suddenly was blurred, you would probably experience that it was easier to read the ambiguous letters if you had read the preceding text. On the extremes of an information-processing continuum between top-down and bottom-up processing, we have remembering and perception of unambiguous, well specified stimuli. Remembering relies heavily on top-down processing, whereas perception of unambiguous, well specified stimuli relies heavily on bottom-up processing. On the other hand, perception of ambiguous or *poorly specified* stimuli may require *both* adequate processing of the present features of the stimuli, and inferences based on top-down information (i.e., an interpretation)[2]. In fact, this thesis is about how perception of poorly specified (i.e., not readily perceived) stimuli can be enhanced by top-down influence, such as from semantic or conceptual information. The poor specification of the stimuli in the studies that the thesis is based on stems from the fact that the speech was presented visually (i.e., without sound), as speech in noise, or as a combination of both, such that varying proportions of the speech signal were difficult to discriminate and identify. That is, elements within the speech signal were difficult or impossible to distinguish. The speech signal under scrutiny in the present thesis can thus be called lacking in linguistic detail, difficult to perceive, ambiguous, unclear, of low resolution, of low intelligibility, as providing paucity in sensory data (Boothroyd, 1988), impoverished or degraded (Luce & Pisoni, 1998), poor (Rönnberg, Samuelsson, & Lyxell, 1998), or poorly specified (Luce & Pisoni, 1998; Rönnberg, 2003*a*, Samuelsson, 1993). Hereby, the term poorly specified is adopted.

## General Problem and Outline

The general problem of this thesis is how information apart from the spoken linguistic signal can be utilised for enhancing perception[3] of the linguistic properties of the signal, when the spoken linguistic signal is poorly specified. The common name for all information that is not constituted by the separate building blocks of the linguistic signal is here specified as *context* (cf. Harley, 1995). The contextual information under study in the thesis consisted of, foremost, *facially displayed emotions*[4] and *topical cues*[5].

The outline of the thesis is as follows. First, the properties and perception of the linguistic speech signal in the auditory, visual, and audiovisual modalities are discussed, in order to establish the characteristics of the signal in each modality. Second, information that may aid perception of poorly specified speech is discussed. This information is conceptually divided into *linguistic*[6], *paralinguistic*, and *prior context*. Third, the individuals' capacity to utilise information from the various sources is discussed, in order to establish the demands the perceiver has to deal with depending on the characteristics of the signal and what other information is available. Fourth, an outline of various effects that facially displayed emotions elicit in perceivers follows, since the major part of this thesis is concerned with effects of facially displayed emotions on speechreading[7] performance. That is, what effects, other than on speech perception, do perceived displayed emotions have, and do these effects interact? Fifth, the objectives of the thesis are stated. Sixth, methodological issues are discussed. Seventh, a summary of the studies follows. Eighth, the conclusions from the results of the studies are presented. Finally, these conclusions provide considerations for studies on perception of poorly specified speech. These considerations are incorporated into a framework for speech perception that comprises specification of the stimuli.

The experiments in this thesis did not follow upon each other such that the outcome of one study determined the hypotheses to be tested in the next study. Rather, the general problem was studied from different perspectives: hypotheses were generated by previous studies, but were not necessarily tested in consecutive order. This approach was chosen because of the ambition to present an overview of all potentially relevant factors in perception of poorly specified speech – the thesis ends up in a framework to this end.

### Integrating Information from Various Sources

To conclude, language perception requires that several sources of information are processed and integrated. This information comes from stimuli as well as from the lexicon, such that the bottom-up information from the stimuli is integrated with and interpreted against top-down information from the lexicon. This processing occurs automatically and effortlessly. However, the information must also be available for conscious manipulation, and the theoretical unit for handling and temporary storage of information is referred to as *working memory*. Working memory has to handle information from stimuli as well as from the lexicon, and also to control attention. In the models of working memory by Baddeley and Hitch (1974) and Baddeley (2000), there is a central unit (i.e., ”the central executive”) that hosts executive attention and uses limited-capacity subsidiary systems for short-term storage of information (e.g., a ″visuospatial sketchpad″, an ″episodic buffer″, and a ″phonological loop″, Baddeley, 2000).

Working memory models account for many processes that a central system has to perform. However, working memory models do not specify where and how integration of bottom-up and top-down information takes place. One reason for this lack of specification is that the integration is automatic, and hence not readily testable[8]. Effects of activation of the lexicon (i.e., top-down influence) have instead been tested by use of linguistic manipulations and priming paradigms (e.g., Luce, Pisoni, & Goldinger, 1990; Marslen-Wilson, Moss, & van Halen, 1996; Samuel, 1981*a*, 1981*b*). Models of language processing, based on results from experiments with linguistic manipulations and priming paradigms, do, on the other hand, neither account for the manipulation of greater amounts of information, such that the capacity of what can be simultaneously stored and manipulated is taxed, nor consider processing of poorly specified stimuli. Therefore, the present thesis ends up in a proposed framework for studies on perception of poorly specified speech that incorporates handling of multiple sources of bottom-up and top-down information. The framework is an attempt to give a comprehensive overview of all major factors of perception of speech, especially poorly specified speech, based on the collective empirical findings from this thesis and other research.

## SPEECH PROCESSING

When the speech signal is poorly specified, sensitivity to the signal and top-down influence are both important for successful speech comprehension. Therefore, properties of the signal, enhancement of the specification of the signal, support for top-down inferences, and the individuals' information processing capacity are discussed.

### The Linguistic Signal

The speech signal in the auditory, visual, and combined (i.e., audiovisual) modalities will be discussed. Depending on whether physical or psychological properties of the speech signal are referred to (i.e., if the speech signal has been sensed by the perceiver or not), different terms will be used. The physical speech signals are referred to as *acoustical* and *optical*, whereas the psychological speech signals are referred to as *auditory* and *visual*, respectively.

*Properties and perception of the auditory speech signal*

The smallest units in language are speech sounds, or phonemes. In spoken language, the phonemes convey the morphemes, which are the smallest meaningful units. When normal-hearing individuals talk face-to-face or over the telephone, listen to someone talking on the radio or on television, they usually have no difficulty hearing what is being said, as long as the speech is loud enough and there is not too much distraction, such as noise. In other words, auditorily presented phonemes are normally easy to perceive, and hence individual differences in listening performance among a normal-hearing population are not readily observable (Lyxell, 1989; Massaro, 1987).

Auditorily perceived phonemes are usually automatically identified. This identification process appears to be categorical – even though the optical properties of speech sounds can be continuous over phoneme borders, each speech sound is automatically *categorised* (but not necessarily *processed*, cf. Massaro, 1998) as a distinct phoneme (Massaro, 1987, 1998; Repp, 1984; Werker & Lalonde, 1988). The perceptual magnet effect (Kuhl, 1991; see also Iverson & Kuhl, 1995, 2000; Walley & Sloane, 2001) suggests that speech sounds are matched against prototypes of phonemes in the lexicon, and categorised on the basis of shortest perceptual distance[9].

In sum, spoken language is based on phonemes, and the phonemes are automatically identified auditorily. The linguistic information within the auditory signal is consequently readily perceived under normal listening conditions.

*Properties and perception of the visual speech signal*

Visual speechreading is speech perception via observation of facial speech movements and without hearing the speech sounds (cf. Lyxell, 1989). In audio-visual speechreading, the individual relies heavily on perception of speech movements but also hears a fraction of the auditory signal (cf. Lyxell, 1989). As we have established, spoken language is based on phonemes. The problem of visual speechreading is that not all phonemes are readily perceived visually.

Even if there is evidence that some individuals are very good at visually identifying speech sounds (e.g., Andersson & Lidestam, 2003; Bernstein, Demorest, & Tucker, 2000), most individuals in a number of empirical studies (e.g., Johansson & Rönnberg, 1995; Lyxell & Rönnberg, 1992; Mogford, 1987; Samuelsson, 1993; Samuelsson & Rönnberg, 1991, 1993) can not visually identify a large enough proportion of phonemes in order to decode words and sentences without support for inferences (i.e., *topical cues*). Furthermore, individuals that are very good at visually perceiving phonemes can be assumed to depend on good viewing conditions, which are not always provided in real life. Consequently, the visual signal alone is insufficient for enabling speech perception and understanding for most individuals in most situations.

Compared to listening to speech under normal conditions, it can be broadly stated that the visual speech signal is less specified – most speech sounds are easy to hear, whereas only a small number of them are correctly identified visually in phoneme identification tasks (e.g., Demorest, Bernstein, & DeHaven, 1996; Lamoré, Huiskamp, van Son, Bosman, & Smoorenburg, 1998; Owens & Blazek, 1985; van Son, Huiskamp, Bosman, & Smoorenburg, 1993; Woodward & Barber, 1960). This lower accuracy in identification has two causes. First, some speech sounds have a place of articulation that is normally hidden from sight. For example, the fricative consonant /h/ does not involve movements by lips or tongue, only an exhalation and (sometimes) the vocal cords are involved. Second, those phonemes that are relatively easy to see are in many instances difficult or impossible to distinguish visually from each other. For example, vocal cord vibration distinguishes between voiced and unvoiced consonants (e.g., /v/ vs. /f/), but is invisible under normal circumstances (Lisker & Abramson, 1964). As a consequence, phonemes that share visual characteristics are called *visemes*. In English, there are about four to six consonant visemes (Summerfield, 1983; Walden, Prosek, Montgomery, Sherr, & Jones, 1977). The number of visemes is associated with differences between talkers and if coarticulation is taken into account (e.g., Auer, Bernstein, Waldstein, & Tucker, 1997). Lidestam, Beskow, and Lyxell (2003) found five consonant visemes for Swedish in one talker.

Compared to reading, the visual speech signal is less specified, and disappears as soon as it has been presented, whereas printed text usually is static and exists optically before the reader even when it has been read (i.e., it is possible to read the text again). In contrast, in speechreading, the stimuli are dynamic ambiguous movements that are presented at a fast pace, and that do not last as optical objects – you cannot make a saccade and look again at the same speech movement. Further, the reader controls the pace when reading the text, whereas the speechreader cannot directly control the speaking rate of the person talking.

In sum, since speech is constituted by phonemes, and since many phonemic features are not readily visually perceivable from speech movements, a considerable part of spoken language is not visually perceived by most individuals. This separates the task of speechreading from listening and reading – in comparison, the visual speech signal is poorly specified. As will be elaborated, successful perception of a poorly specified linguistic speech signal may require cues in addition to the linguistic signal and that the perceiver is skilled in integrating the sources of information. That is, the signal alone is not sufficient if it is too poorly specified.

*Properties and perception of the audiovisual speech signal*

When speech is perceived bimodally, such as audiovisually (with the auditory speech signal being presented in noise or attenuated by sensory hearing loss), and visuo-tactually, perception is often enhanced synergetically. That is, speechreading performance in bimodal conditions is better than the pooled performance from the unimodal conditions (Summerfield, 1983, 1987). This synergy effect may be attributed to three causes. Firstly, that the physical signals from the two modalities complement each other. Secondly, that the signals are fused to a percept that is more than the sum of its parts. Thirdly, a combination of complementarity of signals and fusion.

Summerfield (1983) concluded that hearing impairment and noise produce similar effects on speech perception, since the acoustical information from voicing and manner of articulation typically depends on lower frequencies and prosodic changes, whereas the acoustical information that distinguishes between places of articulation typically consists of spectral detail in the higher frequencies. Noise is more detrimental to high frequency spectral detail than to low frequencies and prosodic changes, which have more energy (Summerfield, 1983). Further, most cases of hearing impairment involve loss of sensitivity at higher frequencies (Lutman, 1983).

As already stated, voicing is an invisible phonetic feature, as it is produced by the vocal cords within the larynx. However, voicing is conveyed by a prosodic change and low-frequency information, and is therefore relatively unaffected by noise and the most common forms of hearing impairment. On the other hand, the acoustical information that distinguishes between different places of articulation is relatively apt to be attenuated by noise, but the place of articulation is in many instances relatively easy to see. In other words, the auditory signal from speech in noise and the visual speech signal from speech movements complement each other, such that phonemes can be identified instead of just low-frequency phonetic features or visemes. In sum, at a physical level, optical and acoustical signals in noise-degraded or higher-frequency attenuated speech complement each other well, such that phonetic features that cannot be heard can be seen, and vice versa (Summerfield, 1983).

At a perceptual level, the auditory and visual signals are integrated, and under certain artificial conditions fused to a percept that is something else than either part. The McGurk effect (McGurk & MacDonald, 1976) is the famous example of how a synchronous presentation of an acoustical and an optical speech signal are fused to a percept that is something else than what was inherent in either signal. For example, an acoustical /mama/ and an optical /tata/ often produces the percept /nana/. The point here is that the synergy effect of audiovisual presentation may in part be explained by the integration – possibly, the integration of two modalities can produce a percept that is more than the sum of its parts. There are also neurophysiological results that suggest that synergy effects of audiovisual presentation may involve neurons that respond stronger to audiovisual stimuli compared to pooled responses of unimodal stimulations (Stein, Wallace, Jiang, Jian, & Vaughn, 1999). Such neurons may be involved in the integration process. A further possibility is that allocation of attention (Samuel, 1990) to stimulus properties in one modality is affected by perceived stimulus properties in another modality, and vice versa, in an interactive fashion. Thereby, perception in both modalities could hypothetically become more sensitive as a function of cross-modal cueing.

Poorly specified auditory, visual, and audiovisual speech were included in this thesis for two main reasons. Firstly, to allow comparisons between the modalities, such that conclusions about whether the effects are modality-specific or amodal may be drawn. Secondly, to manipulate specification of the linguistic signal, above all as a result of the complementarity of natural audiovisual speech compared to unimodal poorly specified speech.

### *Linguistic context*

Phonemes are normally not perceived as separate independent units, but in combinations with other phonemes, such as in a *word* (or lexical) context or in a *sentence* (or syntactic) context. Word context refers to the framing of target phonemes within words versus nonwords; sentence context refers to the framing of target words within sentences. The framing of targets in a linguistic context allows inferences about the target. Linguistic context can also be conceptualised as providing constraints on activation spreading in the lexicon, such as by affecting levels of activation (Marslen-Wilson, 1990). Both word and sentence context are suggested to work top down via the semantic information (i.e., meaning) that the context conveys, and via the formal structure of the language (i.e., grammar), but to affect sensitivity to phonemes in opposite ways (Samuel, 1990).

Word context facilitates detection of phonemes in monitoring tasks, such that phonemes are detected faster and categorically identified when they are presented in words than when they are presented in non-words (Cutler, Mehler, Norris, & Segui, 1987; McClelland & Rumelhart, 1986). Word context also accomplishes phoneme restoration effects (Obusek & Warren, 1973; Warren, 1970; Warren & Obusek, 1971; Warren & Warren, 1970) and reduces sensitivity to detect that phonemes are replaced by noise as opposed to masked by noise, suggestively by gearing on-line inferences toward the phoneme that is consistent with the word (Samuel, 1990). To conclude, word context seems to increase phoneme identification speed, but at the expense of sensitivity to phonemic detail (i.e., phonemes in noise; sublexical information, Samuel, 1990).

Sentence context, on the other hand, can increase sensitivity to phonemic information (Samuel, 1981*a*), suggestively by facilitating focus of attention to the relevant locations within lexical representations, such that a comparison between the stimulus and the representation is facilitated (Samuel, 1990). However, there are still phonemic restoration effects within sentence context (Samuel, 1981*a*), and the constraints by sentence context can also affect identification of an attenuated word (Isenberg, Walker, & Ryder, 1980). The bias to report that no phoneme is missing can be suggested to depend on sentence context facilitating word recognition (i.e., lexical identification, identification of a target stimulus as a word), such that on-line, perceptual, inferences are made, with decreased sensitivity to phonemic information as a result. That is, sentence context effects have word context effects as a byproduct. Samuel (1990) concluded that lexical identification affects perception top down and directly, whereas higher-level knowledge (i.e., sentence context) can affect postperceptual processing, and that attentional allocation determines perceptual processing. Boothroyd (1988) concluded that lexical, syntactic, topical, and semantic context in the message enhance speechreading performance, and that none of the effects is sufficient, but that they work synergetically in combination.

### Paralinguistic Context: Cues that Accompany the Linguistic Signal

#### *Auditory paralanguage*

The term paralanguage is usually delimited to auditory speech (e.g., Crystal, 1997; Knapp & Hall, 1997; Trager, 1958). Paralanguage is specified as *how* something is said, not *what* is said (Knapp & Hall, 1997), as accompanying the linguistic information, and as expressed with vocal sounds (Crystal, 1997). For example, vocal sounds can convey happiness with phonemes and laughter (Trager, 1958). Other examples of paralanguage are speech modifiers such as pitch, rhythm, intensity, and pauses; as well as vocalisations such as crying, groaning, sneezing, snoring, and yawning (Trager, 1958). Common to all these variants of paralanguage is that they can convey meaning. For example, yawns can signal boredom, and high pitch can be a sign of aggression.

How the intended meaning of an utterance is interpreted does not only depend on the linguistic information of the speech signal, it also depends on the situational context and the paralanguage. That is, an utterance has a ”core” linguistic meaning as well as a pragmatic meaning (Crystal, 1997). Paralanguage can also interact with the linguistic meaning to modify the perceived pragmatic meaning of the utterance. For example, the linguistic meaning can be emphasised (e.g., ”I'm happy” with a smile) or contradicted (e.g., ”I'm happy” with a frown), resulting in the utterance probably being interpreted very differently.

*Visual paralanguage*

As already stated, the term paralanguage is usually delimited to auditory speech. If Knapp and Hall’s (1997) definition of paralanguage is accepted, however, the visual speech signal can also be claimed to have paralanguage. From the definition ”not what is said, but how it is said” follows that the paralanguage is not restricted to the speech movements, rather, all visual cues that the talker conveys and that can influence how an utterance is perceived and interpreted are included in visual paralanguage. For example, intonational information is primarily obtained from the upper parts of the face (i.e., forehead and eyes) in visual speech perception (Lansing & McConkie, 1999).

Consequently, under Knapp and Hall's (1997) definition of paralanguage, displayed emotions, gestures, gaze, body movements, posture, and interpersonal distance are included. Some of the sources of paralanguage convey emotion, and some do not. The forms of paralanguage that convey emotions will from this point on be denoted *displayed emotions*, and the forms that do not convey emotions will be denoted *gestures*. Further, some of these sources of paralanguage are conveyed by the face, whereas some paralinguistic signals are conveyed by other parts of the body. The four instances of visual paralanguage in combinations of contents and location will therefore be discussed next. Note that a behaviour may be sorted under both displayed emotions and gestures, since many gestures derive from emotional expressions. Two of these instances have been empirically proven to affect visual speechreading performance (and are marked with asterisks, see Figure 1). No reports of empirical tests of effects on accuracy in perception of poorly specified speech have been found in the literature for the other two instances.

*Facially displayed emotions.* The conveyance of emotion when something is said can radically change the perceived meaning of the utterance. For example, if someone says ”I'm sorry that you lost” to someone who has just lost a tennis match, and looks happy, this will indeed be interpreted differently compared to if the person who uttered the phrase looked compassionately sad. In the former case, the utterance would probably be interpreted as ironic or malicious, that is, such that the person actually was happy that the other one had lost. According to the theoretical framework of Bruce and Young (1986), facial expressions are processed in parallel with visual speech. This parallel processing was hypothesised to occur in separate input modules (cf. Fodor, 1983), which suggests that perception of facial expressions may not be able to facilitate perception of visual speech (i.e., speechreading). However, a facilitatory effect of happy but not of sad facial expression on speechreading performance has been found (Johansson, 1997; Johansson & Rönnberg, 1996). The conclusion made by Johansson (1997, 1998) was that positive displayed emotion affects the state of the perceiver, such that approach behaviour (Davidson, Ekman, Saron, Senulis, & Friesen, 1990) is elicited, increasing motivation to focus on the talker's speech. However, facially displayed emotions may not merely affect the state of the perceiver: specification of the signal (i.e., articulatory distinctiveness) may also be enhanced, and the fact that the displayed emotions convey the emotional meaning of the utterance may facilitate lexical access or support inferences. One way to test if articulatory distinctiveness is affected by displayed emotions is to keep articulation constant while displaying emotions, by presenting the stimuli by means of a synthetic talking head[10]. Another way is to test whether identification of phonemes out of linguistic context is affected by displayed emotions. In order to test if displayed emotions facilitate lexical access or support inferences, illustrating displayed emotions may be compared with cue-words that convey the emotional meaning. Motivation to focus on the talker's speech as a function of displayed emotion may be assessed by means of questionnaires.

*Extra-facially displayed emotions*. Emotion can be displayed with body movements and posture as well as with facial expressions (e.g., Lemeignan, Aguilera-Torres, & Bloch, 1992; Montepare, Goldstein, & Clausen, 1987; Sogon & Matsutani, 1989; Wallbott, 1998). For example, depression and sadness are associated with drooping posture and a tendency to look down (Argyle, 1988), while anger is associated with larger and faster movements (Montepare et al., 1987; Wallbott, 1998). Effects of extra-facially displayed emotions on accuracy in perception of poorly specified speech have not been found in the literature.

*Extra-facial gestures*. Gestures can further be divided into pantomimic and non-pantomimic gestures (Cohen & Otterbein, 1992), emblems (Johnson, Ekman, & Friesen, 1975), and illustrators (Ekman & Friesen, 1972). Pantomimic gestures do not require speech to convey meaning and can thus be understood in the absence of speech, whereas non-pantomimic gestures can merely moderate the meaning of speech and thus not be understood without a speech signal (Cohen & Otterbein, 1992). Emblems are pantomimic in the sense that they can substitute speech – they are nonverbal acts that are clearly understood by the majority of members of a culture (Johnson et al., 1975). Illustrators can be both pantomimic and non-pantomimic – they illustrate what is spoken, but some illustrating gestures can be understood without speech, for example, ″up″, ″me″, and ″there″. Extra-facial gestures can enhance visual speechreading performance when they are provided as correct cues (Berger & Popelka, 1971; Popelka, Lexington, & Berger, 1971). They can also impair visual speechreading performance when presented as false cues (Popelka et al., 1971). These results conclusively demonstrate that paralanguage can facilitate speech perception without enhancement of the specification of the linguistic signal – as the gestures were extra-facial, they cannot affect speech movements in the face. Further, because the gestures and sentences in Popelka et al. were not emotional, the result can be taken as support for the idea that paralanguage can enhance speech perception without a mediating effect of emotional state in the perceiver.

*Facial gestures*. The non-emotional facial expressions can be categorised as gestures, and can as such be pantomimic, non-pantomimic, emblems, and illustrators. Sticking out the tongue can be a pantomimic gesture of dislike or teasing as well as an emblem (i.e., a greeting signal in Tibet, Axtell, 1998; Hergé, 1962). Gaze can be utilised as illustrators by, for example, drawing attention to objects in the environment. However, effects of facial gestures on perception of poorly specified speech have not been found in the literature.

In sum, facially displayed emotions and extra-facial gestures can enhance visual speechreading performance when they illustrate the linguistic meaning of an utterance. As will be elaborated in this thesis, extra-facially displayed emotions and facial gestures can be hypothesised to follow the same pattern – they are most likely to improve perception of poorly specified speech, as long as they provide correct and meaningful cues to some aspect of the meaning of an utterance.

| | Contents | |
|---|---|---|
| Location | Displayed emotions | Gestures |
| Facial | Facially displayed emotions* | Facial gestures |
| Extra-facial | Extra-facially displayed emotions | Extra-facial gestures* |

*Figure 1*. Taxonomy of visual paralanguage.

**Prior Context: Priming the Lexicon for the Signal**

Utterances are seldom made in an informational vacuum. For example, when an utterance is made in everyday communication, on the TV news, or in a narrative, such as in a written text, or a fairy-tale, we usually know who is talking. We also usually know the purpose of the person talking, and what he or she is talking about. In short, we know what is likely to occur and to be said – we have *topical cues* to relate the utterances to (cf. prior context, Harley, 1995).

If a topical cue is to be useful in perception of poorly specified speech, such as in speechreading, it has to be presented before or during, but not after, the stimulus (e.g., Garstecki, 1976; Samuelsson, 1993). That is, it has to serve the function of a *prime*; to precede the signal such that access to the lexicon is facilitated when the signal subsequently is to be identified, by initiating activation spreading in the lexicon (e.g., Marslen-Wilson, 1990). In order for the spreading activation to be facilitatory, the prime has to contain relevant cues about some property of the utterance, such as an accurate description of the topic or script (Abelson, 1981; Marslen-Wilson et al., 1996; Schank & Abelson, 1977). The facilitation is conceptualised as constraints on the spreading activation (e.g., Bock & Levelt, 1994).

Most speechreaders need prior contextual cues, such as topical cues, in order to get visual speechreading scores above levels produced by pure guessing on open-ended responses (e.g., Garstecki & O'Neill, 1980; Samuelsson & Rönnberg, 1991; Smith & Kitchen, 1972). Prior contextual cues also seem to be required in order for most skilled speechreaders to be able to understand the major part of visual speech (e.g., Lyxell, 1994; Rönnberg, 1993; Rönnberg et al, 1999). The facilitatory effect of prior contextual cues is not restricted to visual speechreading. For example, it also exists in reading (e.g., Neely, 1977, 1991). Lansing and Helgeson (1995) suggested that semantic priming (cf. prior contextual cues) serves the same function in visual spoken word recognition, visual text recognition, and auditory word recognition – by activating the lexicon.

*Cue distinctiveness* refers to how many items in the lexicon that share the same cue (cf. Hunt & Smith, 1996). How efficiently the lexicon is activated by cues is determined by cue distinctiveness with regard to cued recall in memory research (cf. Hunt & Einstein, 1981; Hunt & Smith, 1996). It has also been demonstrated that cue distinctiveness affects speechreading performance (Samuelsson, 1993; Samuelsson & Rönnberg, 1991, 1993). It can be concluded that word recognition (i.e., lexical identification) in speechreading is accomplished by the successful combination of semantic priming due to topical cues, and perceived visual linguistic information (cf. Samuelsson, 1993).

### Interim Summary

This far we have looked at the speech signal as it consists of phonemes and their combination into a linguistic context; paralinguistic context; and prior context. Specification of the speech signal has been considered, and how it is affected by modality. Linguistic, paralinguistic, and prior context all affect perception of poorly specified speech. Linguistic and prior context can be assumed to affect speech perception via their semantic properties, that is, by conveying meaning. For example, it may be easier to perceive the word ″advantage″ if we know that it is uttered by a tennis umpire or if we hear it in a sentence such as ″the new method has a clear advantage over the old one″. Paralanguage may convey semantic cues, but can also affect pragmatics. For example, the talker's emotional state may be inferred.

It can be generally stated that both linguistic and prior context reduce the number of possible candidates for a lexical match, and this facilitates phoneme perception (Cutler et al., 1987; Marslen-Wilson et al., 1996; McClelland & Elman, 1986; Samuelsson, 1993). This facilitation may be due to enhanced sensitivity as well as to postperceptual inferences. In fact, there has been a debate on the existence of sensitivity effects and response bias (e.g., Farah, 1989; Massaro & Oden, 1995; McClelland & Elman, 1986; Pitt, 1995*a*, 1995*b*; Rhodes, Parkin, & Tremewan, 1993; Rhodes & Tremewan, 1993), but this debate lies beyond the scope of this thesis.

### Information-Processing Capacity and Speech Processing

We have concluded that there are different sorts of information that may have to be processed within speech perception. We have also briefly discussed *how* the different sorts of information are processed. Next we will discuss what *demands* the perceiver encounters when processing a poorly specified speech signal, paralanguage, and semantic cues. That is, what capacities are required in order for the poorly specified speech to be efficiently perceived?

*Bottom-up processing capacity*

Obviously, the less specified a stimulus is, the more important it becomes that the fraction of the stimulus that *is* perceived is perceived correctly – there is no large margin for error. As an analogy, the less money you have, the more important it becomes that it is not wasted. If we compare visual speech to auditory speech, auditory speech is better specified, and contains more linguistic information. If a phoneme or features of phonemes are missing in audition, speech perception will not be greatly affected. In fact, missing phonetic information may not even be detected as a consequence of the phoneme restoration effect (cf. Samuel, 1981*a*, 1981*b*, 1991, 1997; Warren & Warren, 1970). On the other hand, if a viseme is misperceived in visual speechreading, the effect will often be detrimental to speech perception, since so few pieces of information are conveyed even under good circumstances.

Further, the optical speech signal does not endure, and the information is presented at a fast pace. This means that not only do the pieces of information need to be correctly identified, they also have to be processed in a fast and automatic manner. Conclusively, the characteristics of a poorly specified speech signal (such as visual speech) require the perceiver to be accurate and fast. In other words, bottom-up skills are basic requirements for perception of poorly specified speech.

In accordance with these theoretical assumptions, empirical findings have shown that lexical decision-making speed (Lyxell & Holmberg, 2000), phoneme identification (Bernstein et al., 2000), and word decoding[11] without topical cues (Lyxell & Holmberg, 2000), have shown to be associated with visual speech-reading performance in group data on normal-hearing subjects. Group data on hearing impaired and deafened individuals have revealed significant correlations between visual speechreading performance and phoneme identification (Bernstein et al., 2000; Demorest et al., 1996); word decoding without topical cues (Andersson, Lyxell, Rönnberg, & Spens, 2001; Lyxell & Holmberg, 2000); and visual–neural speed (Rönnberg, Arlinger, Lyxell, & Kinnefors, 1989). Andersson and Lidestam (2003) reported a case study demonstrating that superior bottom-up processing abilities (e.g., phoneme identification) can be a base for speechreading of more linguistically complex utterances (i.e., sentences).

To conclude, bottom-up skills that enable fast and accurate processing are *necessary* for perception of a volatile and poorly specified speech signal, since they provide the information from the stimuli. If the signal is poorly specified, as much information as possible must be extracted, and the extracted information must not be distorted.

*Top-down processing capacity*

Fast and accurate information processing is necessary but not *sufficient* for successful perception of a volatile and poorly specified speech signal. If the signal is poor, inferences must be made about the information that is missing within the signal. These inferences are, as we have established, executed top down: information from the lexicon is matched and integrated with the bottom-up information. As a consequence, the poorer the specification of the signal, the greater the room for inferences becomes. That is, more inferences have to be made, and the more important it becomes that the inferences are adequate.

The empirical findings with regard to associations between speechreading performance and top-down skills are also in accordance with this notion. Working memory capacity, as measured by the reading span task (cf. Baddeley, Logie, Nimmo-Smith, & Brereton, 1985), and ability to infer missing information in various verbal cloze tests (Bode, Nerbonne, & Sahlstrom, 1970; Lyxell & Rönnberg, 1989; Sanders & Coscarelli, 1970; Williams, 1982), are correlated with visual speechreading performance for normal-hearing individuals in group studies (Lyxell & Holmberg, 2000; Lyxell & Rönnberg, 1987*a*, 1989). Group studies on hearing impaired individuals have produced similar results (Lyxell & Holmberg, 2000; Lyxell & Rönnberg, 1989). Case studies on hearing impaired and deaf speechreading experts have corroborated the conclusion that top-down skills can be the basis for excellent speechreading skill (Lyxell, 1994; Rönnberg, 1993; Rönnberg et al, 1999). Rönnberg, Samuelsson, et al. (1998) put forward a hypothesis stating that superior speechreading skill only can be achieved if top-down processing skill, especially working memory capacity, surpasses a hypothetical threshold level. There are, however, reasons to dispute this hypothesis. For example, skilled speechreading may be based on superior bottom-up skills (Andersson & Lidestam, 2003).

Just and Carpenter (1992) demonstrated that working memory, as measured by the reading span task, mediated reading comprehension, such that only participants with high reading span had resources to take meaning into account when parsing. Visual speech processing ability is also associated with ability to use linguistic constraints (Boothroyd, 1988). Reading span performance is also correlated with speechreading performance (Lyxell & Holmberg, 2000). Together these results suggest that some aspect of executive function or working memory function can mediate how linguistic information and semantic cues are handled and perceived.

### *The combination of bottom-up and top-down information*

The collective empirical picture thus concurs with theory in that both bottom-up and top-down skills are required for successful visual speechreading. According to Baddeley's (2000) model of working memory, the central executive is the unit that integrates these sources of information. The capacity of the central executive can be assumed to be taxed in the reading span task (Baddeley et al., 1985; Rönnberg et al., 1989), and therefore the significant correlation between performance on the reading span task and semantically (topically) cued visual speechreading performance (Lyxell & Holmberg, 2000) may be interpreted as support for the assumption that the central executive integrates top-down and bottom-up information. However, there is a problem in that the reading span task is complex and involves many types of information processing. For example, *attention* is crucial both to solving of the reading span task and to speechreading. Attention is a very important aspect of the central executive (e.g., Baddeley, 2000; Engle, 2002; Kane & Engle, 2002, 2003; see also Richardson et al., 1996, for an overview of working memory models; and Gathercole & Baddeley, 1993, for an overview of research on the role of working memory in language). Attention has further been suggested to explain variance in sentence-based speechreading performance together with short-term visual memory, and processing speed (i.e., visual attention, Boothroyd, 1988). Hence, it is difficult to draw inferences about which of general working memory capacity, capacity of its parts (such as the central executive), or certain functional aspects of it (such as attention) is most essential for the capacity to perceive poorly specified speech.

Which of bottom-up and top-down skill is the most dominant factor in perception of visual speech can be assumed to depend on task characteristics. The more information to temporarily store and manipulate, and the more information that is missing within the speech signal, the greater will be the relative importance of top-down processing. If the task merely requires identification of visual features, such as identifying a phoneme, there is not much top-down processing involved, and top-down skill will thus not be taxed. On the other hand, if the task is to perceive a stimulus that is linguistically more complex and poorly specified, such as in speechreading a sentence visually, there may be many missing pieces of information to infer, and a lot of information from both the signal and the lexicon to simultaneously manipulate and store.

*Sensory impairment and information-processing capacity*

Does sensory impairment result in sensory, perceptual, or cognitive compensation? More specifically, does hearing impairment and deafness result in better speechreading ability? The answer seems to be that it usually does not, but some elaboration on this statement is required.

The conclusion that speechreading ability is *not* usually improved as a consequence of hearing loss or deafness comes from a number of group studies that have compared mean speechreading performance in hearing impaired, deaf, and normal-hearing populations (e.g., Clouser, 1976, 1977; Conrad, 1977; Erber, 1972; Hygge, Rönnberg, Larsby, & Arlinger, 1992; Lyxell & Rönnberg, 1987*b*; Lyxell & Rönnberg, 1989; Mogford, 1987; Rönnberg, 1990; Rönnberg, Öhngren, & Nilsson, 1982, 1983). Thus, there seems to be no spontaneous perceptual compensation for hearing impairment in the majority of hearing impaired and deafened individuals. Correlations between length of time with hearing loss or deafness and speechreading performance, and between degree of hearing loss and speechreading performance have generally not been reported to be significant (Erber, 1972; Lyxell & Rönnberg, 1987*b*, Tillberg, Rönnberg, Svärd, & Ahlner, 1996). Sensory, perceptual, or cognitive compensation as a general rule in hearing impairment or deafness would be expected to be reflected in improved mean speechreading performance. As the aforementioned results suggest, however, this does not seem to be the case.

There are, however, also results that *do* indicate that speechreading performance is associated with hearing impairment and hearing loss. Superior speechreading skill is found mainly in hearing impaired or deaf individuals – reported cases with superior speechreading skill have in common that they are hearing impaired or deaf (e.g., Andersson & Lidestam, 2003; Lyxell, 1994; Rönnberg, 1993; Rönnberg et al, 1999). Group studies have also demonstrated that superior speechreading ability is much more frequent among the hearing impaired and deaf than among the normal hearing (e.g., Bernstein et al., 2000; Bernstein, Demorest, Coulter, & O'Connell, 1991; Demorest & Bernstein, 1997). One factor that has been suggested to explain that speechreading performance in the deaf and hearing impaired in general is *not* superior to that of the normal hearing is less exposure to speech, such that words (as lexical items) are less familiar to the hearing impaired and deaf than to the normal hearing (Auer, Bernstein, & Tucker, 2000).

Phonology, the psychological representation of speech sounds, is an important factor for the interpretation of lip movements as meaningful units (i.e., morphemes). The logic is as follows: speech is based on phonology (e.g., Liberman, 1996), and if the phonemic configuration cannot be recognised, speech perception is difficult. If lip movements cannot be associated with phonemes, speechreading becomes difficult. Deterioration of phonological representations should therefore be detrimental to speech perception, including visual speech perception. Phonological representations deteriorate as a function of time in adults with acquired hearing loss (Andersson, 2002; Andersson & Lyxell, 1998), and there is a negative correlation between speechreading ability and onset of deafness (Berger, 1972). Conrad (1979) claimed that there is no language basis to build on for learning speechreading if deafness occurs prior to language acquisition. In line with these assumptions, Andersson et al. (2001) found that phonological awareness is correlated with speechreading performance. Consequently, hearing impairment can *impede* speechreading via deterioration of phonological representations.

Variance in speechreading performance is thus larger within populations of hearing impaired and deaf than among the normal hearing – a small proportion develop superior speechreading skill, whereas others lose the prerequisites for speechreading. In other words, deafness and hearing impairment seem to have a Matthew effect (XXV:29, as cited in Stanovich, 1986) in speechreading. It can be hypothesised that compensation for hearing loss or impairment by skilled speechreading is facilitated if an individual has good enough skills for speechreading before the onset of the hearing loss or impairment. That is, if the individual is able to visually extract a large enough proportion of the phonetic information as well as to infer a large enough proportion of the information that could not be extracted, speechreading is possible. If the hearing impaired or deaf individual is successful in perceiving the visual speech often enough, the result can be fine-tuning of inferences and phoneme identification, as well as practice in focusing attention, prioritising information, and logistical handling of input and lexical information. In other words, provided that the individual manages to speechread well enough, he or she will be able to calibrate various aspects of speech perception and practise aspects of executive function or working memory. This can only be possible if the percepts can be proven valid often enough (i.e., that the speech is understood). Without some accuracy in speech perception, calibration cannot occur, nor will there be intrinsic reinforcement (cf. Skinner, 1971) for speechreading. It can be hypothesised that speech perception must be above some threshold level in order for calibration of phoneme identification to be possible. Perception can be drastically improved if topical cues are provided, allowing more top-down influence.

Age of onset of hearing impairment or loss is suggested to be another important factor for development of speechreading skill. If hearing impairment or loss occurs after development of spoken language skills but relatively early in life, the prognosis for reaching high speechreading ability is improved (cf. Andersson & Lidestam, 2003; Rönnberg, 1993; Rönnberg, Andersson, et al., 1998; Rönnberg, Samuelsson, et al., 1998). Differences in age of onset of hearing impairment or deafness *within* the age span may also account for qualitative differences in speech processing (cf. Andersson & Lidestam, 2003; Boothroyd, 1988). It has been suggested that cortical plasticity of a developing brain can make reallocations of processing capacity possible to allow visual speech input to be processed in cortical areas that used to process auditory speech information (e.g., Rönnberg, 2003*b*; Rönnberg, Andersson, et al., 1998; Rönnberg et al., 1999). This would ensure phonological processing, and in a situation when it is necessary to perceive speech but there is no auditory information to process, calibration of visual phonemic perception then becomes both necessary *and possible*. The result would then be that the visual speech input is so specified that the task can be solved much more bottom up than is the case for those who cannot identify as much phonemic information in the optical speech.

## EFFECTS OF FACIALLY DISPLAYED EMOTION ON THE PERCEIVER

Facially displayed emotions are universally recognised as six to ten more or less distinct basic categories (e.g., Ekman; Fridlund; and Izard; as cited in Cornelius, 2000). Explanations for the aetiology of facially displayed emotions range from evolutionary (e.g., Ekman et al., 1987) via cognitive (e.g., Arnold, 1960; Frijda, 1986) to social constructivist (e.g., Averill, 1980; Harré, 1986; Oatley, 1993), according to Cornelius (2000). Fridlund (1991, 1994), and Fridlund and Gilbert (1985) claimed that the paralinguistic function of displayed emotion is essential. This implies a cognitive perspective, as appraisal (Arnold, 1960) of the emotion can be assumed to be involved.

### Processing of Emotional Cues

When facially displayed emotions are perceived, many reactions within the perceiver are elicited. For example, the displayed emotions are automatically and instantaneously reflected by activity in the facial muscles in the perceiver (Dimberg, 1997; Dimberg & Öhman, 1996; Esteves, Dimberg, & Öhman, 1994; Lundqvist & Dimberg, 1995). Arousal (e.g., Cowie, 2000; Davidson et al., 1990;

Dimberg & Öhman, 1996) and attitude (Cowie, 2000) are also affected. These reactions are mediated by many factors, for example spatial attention (Holmes, Vuilleumier, & Eimer, 2003), context of other displayed emotions (Russel & Fehr, 1987), and visual field to which the stimuli are presented (Davidson, Mednick, Moss, Saron, & Schaffer, 1987). The role of displayed emotions is suggested to have been established relatively early in the evolution, as also primates are sensitive to them (e.g., Ghazanfar & Logothetis, 2003). Hence, facially displayed emotions affect the perceiver in many more ways than those studied within the scope of this thesis.

Regardless of how emotional information is presented (i.e., facially, prosodically, or linguistically), it is suggestively handled by the same general processor (Borod et al., 2000). Neuroscientific evidence suggests that production and perception of displayed emotions are processes that are executed in the same systems as cognitive processes, and that emotion and cognition hence should not be divorced (Erickson & Schulkin, 2003). There is even evidence from processing on the *neuronal* level that concurs with this reasoning: Sugase, Yamane, Ueno, and Kawano (1999) reported that there are face-selective neurons that encode presence of a face first, and emotional information thereafter. In a recent overview of the research on verbal and nonverbal communication, Jones and LeBaron (2002) argued that verbal and nonverbal messages should be studied as inseparable phenomena when they occur together.

The combined empirical picture thus suggests the facilitatory effect on speech perception to be the same whether the emotional meaning of an utterance is conveyed by displayed emotions or by verbal cues. Also, emotional state in the perceiver can be assumed to be affected not only by displayed emotions, but also by other stimuli.

### Emotion-Related State in the Perceiver

The states in the perceiver that are affected by displayed emotions include arousal (cf. Cowie, 2000; Davidson et al., 1990; Dimberg & Öhman, 1996) and attitude (Cowie, 2000). There appears to be an interaction between emotion-related states in the perceiver and language processing. Johansson (1997, 1998) suggested that happy facial displays facilitate visual speechreading due to approach behaviour (Davidson et al., 1990), since she obtained facilitation from positive, but not negative, displayed emotion. This notion was, however, not supported in Lidestam et al. (2003), where no correlation between ratings of motivation (as an indicator of approach behaviour) and performance in speech perception tasks was found. Further, Niedenthal, Halberstadt, and Setterlund (1997) found that musically induced emotional state of the perceiver is associated with facilitation of response to words categorically related to that emotion (e.g., happiness and a word that is associated with happiness). However, the facilitation requires a categorical relationship, as relationship within valence[12] (e.g., happiness and a word associated with love, and not with happiness) did not suffice for obtaining facilitation. This may be interpreted such that cues to valence (that comprises various emotions) are too general as cues, whereas cues to the emotional category are distinct enough as cues for facilitation of retreival (cf. Hunt & Einstein, 1981; Hunt & Smith, 1996). Also, situational context mediates ratings of affect in auditory speech (Cauldwell, 2000), and as will be elaborated within this thesis, this finding is hypothetically valid across modalities.

In sum, displayed emotions elicit various responses within the perceiver, and emotion-related states in the perceiver can affect language processing. Motivation is an emotion-related state that has been suggested to facilitate speechreading performance, based on the assumptions that only happy displayed emotion facilitates speechreading performance, and that approach behaviour (Davidson et al., 1990) is involved (Johansson, 1997, 1998).

## OBJECTIVES

The general aim of this thesis was to study how different sources of information beside the spoken linguistic information are integrated and utilised in speech perception. To examine all available sources of information that can exist beside the linguistic information of all possible forms of speech is not possible within the limits of a doctoral thesis; thus, certain delimitations were made. Firstly, the speech signal was confined to be poorly specified, in order to allow paralanguage and semantic cues to distinctively facilitate speech perception. Visual presentation of speech was used in all studies (I–V). When auditory presentation of speech was used unimodally and bimodally (Studies III & V), it was degraded by a broadband ″white″ noise (cf. Noble & Perrett, 2002). Secondly, the examined sources of information beside the spoken linguistic information were delimited to facially displayed emotions (Studies I, II, IV, & V); cue-words for topical context (i.e., topic, Study III; script, Study V); and cue-words for emotional content (Study V).

The specific purposes of the thesis were as follows. Firstly, to examine if facially displayed emotions can facilitate visual and audiovisual speechreading. Studies I, II, IV, and V tested whether facially displayed emotions can improve visual speechreading performance when they illustrate strong positive and negative valence of utterances, that is, when language and paralanguage are in consonance.

Secondly, to study how this facilitation is accomplished. Three rivalling hypotheses were tested. The hypothesis that displayed emotion affects visual speechreading via emotion-related state in the perceiver (i.e., motivation) was tested in Study IV. This was accomplished by testing effects of displayed emotion and linguistic specification on person impression, attitude to the task, and speechreading performance, and correlating these measures. The hypothesis that perception is enhanced as a function of improved articulatory distinctiveness was tested in Study V, by use of a synthetic talking head (Beskow, 1997) that can display emotions without affecting the articulatory movements (Beskow, 1995), and by measuring effects on phoneme identification in a nonsense syllable task. The hypothesis that displayed emotions can facilitate perception of poorly specified speech by conveying semantic cues and thereby providing constraints on activation spreading in the lexicon was tested against the articulatory distinctiveness hypothesis in Study V, by presenting emotional cues as text and as facially displayed emotions.

Thirdly, to find information-processing correlates to processing of poorly specified speech, in order to contribute to the mapping of how bottom-up and top-down processing is utilised in processing of poorly specified speech. The association between bottom-up skills and perception of poorly specified speech was tested in correlational designs in Studies I (i.e., the visual word discrimination and word recognition tasks vs. the visual speechreading task), II (i.e., visual lexical decision-making speed and accuracy vs. visual speechreading), and V (i.e., phoneme identification vs. visual speechreading in the auditory, visual, and audiovisual modalities). The contribution of top-down processing to perception of poorly specified speech was examined in all studies within this thesis by testing the effects of topical cues (as prior context), emotional cues (as paralinguistic context), or both in combination. The correlation between heavy top-down processing and speechreading was further tested in Study II (i.e., reading span vs. visual speechreading).

Fourthly, to present a comprehensive conceptual framework for speech perception (including perception of poorly specified speech), incorporating semantic priming, the linguistic and paralinguistic signals, and differences in processing capacity. In addition to the tests presented under the first three purposes, the following was also studied. Studies III and V tested the effects of prior context on listening to speech in noise, visual, and audiovisual speechreading. Further, speech perception in the auditory, visual, and audiovisual modalities was compared with regard to effects of topical cues (as prior context, Studies III & V) and emotional cues (as paralinguistic context, Study V). Specification of the linguistic signal was varied by modalities and type of talker[13] (Studies III & V), and by manipulation into a lower and higher specification list (Study IV). Specification of the paralinguistic signal was studied by including control measures of perceived valence and strength of facially displayed emotions (Studies II, IV, & V). Individual differences in processing capacity in relation to perception of poorly specified speech was further studied by testing the interaction between skill level and facially displayed emotion in Study I.

## METHODOLOGICAL ISSUES

Valid claims of empirical knowledge require that theory of science, methods and the phenomena under study match. The extent to which the claims of knowledge are valid and generalisable also depends on methodological issues. Considerations regarding theory of science, methods, and rationales for the methodological choices are therefore presented.

### Approaches and Considerations due to Theory of Science

A first distinction of this scientific endeavour is that it is based on empirical results. This means that the knowledge is derived from observations – that knowledge about the objects under study cannot be derived from mere application of rules, as in formal sciences (i.e., mathematics and philosophy). Secondly, a quantitative approach was used, meaning that the purpose was to obtain empirical results that can be generalised to populations by drawing inferences from samples. Further, quantitative measures of, for example, effects and effect sizes were obtained by the use of statistics. The overall aim in interpreting the statistical results was to understand the set of data at hand, following Tukey (1977). The fact that one pointwise significant effect along a continuum is likely to be due to a Type 1 error, and that a uniform pattern of results is likely to indicate a true effect, was considered. That is, not only statistical significance, but also patterns in the data were considered.

In order to allow valid conclusions from the combined data, Platt's (1964) strong-inference strategy was adopted, meaning that multiple hypotheses on a particular phenomenon are tested, as compared to just one in the falsificationist framework (Massaro, 1987; Popper, 1959). When different sets of data and tests produce the same type of results, we are less likely to be wrong (Garner, Hake, & Eriksen, 1956; Hammersley & Atkinson, 1995). In the case of remaining rivalling hypotheses, the principle of Ockham's razor should be applied – the less complicated explanation tends to be the one that is closest to the truth.

In the studies, three rivalling hypotheses for one phenomenon were tested, and converging results from different tests were sought.

## Methods

*Participants*

A total of 399 normal-hearing adults with good visual acuity participated for a total of approximately 300 man-hours. Generalisations of the results to deaf and hearing impaired populations cannot be made without some caution. Although group data suggest that normal hearing, hearing impaired, and post-lingually deaf populations do not differ with respect to important aspects of visual speech perception (e.g., mean speechreading performance, Lyxell & Rönnberg, 1989; Rönnberg, 1990), the results from this thesis cannot uncritically be regarded as valid for other populations than normal-hearing adults.

*Designs*

Experimental designs were used for studying how facially displayed emotion and topical cues are used in speech processing, in order to allow valid inferences about causality. In many instances, the designs were factorial for the purpose of allowing interactions between the variables to be observed. Correlations were calculated for studying the involvement of bottom-up and top-down components in perception of poorly specified speech.

*Experimental paradigms and tests*

A perceptual identification paradigm (cf. Lively, Pisoni, & Goldinger, 1994) was used for the sentence-based speechreading task, the word decoding task (denoted the sentence identification task and the word identification task, respectively, in Study V), and the phoneme identification task in the studies. A visual matching paradigm was used for the word discrimination task, and a visual–lexical matching paradigm was used for the word recognition task in Study I. A reading span test (Baddeley et al., 1985; Rönnberg et al, 1989), and a lexical decision-making test (Rönnberg et al.) were used in Study II. Two questionnaires measured attitude to the task and perceived extraversion (after Asch, 1946) in Study IV. Reliability of the tests in Study IV was reported by means of Cronbach's alpha coefficient.

*Stimuli*

The optical speech stimuli consisted of speech from video-recordings of four human[14] male actors and two synthetic male-like talking heads. The acoustic speech stimuli consisted of natural speech from the soundtracks of recordings of two of the human actors, and was masked by broadband ”white” noise, in order to avoid threshold effects in the auditory conditions and ceiling effects in the audiovisual conditions. The mean signal-to-noise relationship was approximately –10 dB (A)[15] in the speechreading tasks, with a mean total sound pressure level of about 73 dB (A) in Study III, and about 65 dB (A) in Study V. The stimulus lists were validated for typicality and valence by a total of 31 raters. The stimuli were validated with regard to distinctiveness of displayed emotion as a within-group control in Studies II and IV, and by 15 independent raters in Study V.

*Operationalistic issues*

Accuracy in speech perception, measured as the proportion of phonemes that were correctly rendered back was used as dependent variable for the word and sentence identification tasks (with the exception of Study I, where proportions of correctly rendered word stems were scored). That is, the measure was the rendered linguistic properties of the utterances, not gist nor classification. This operationalisation was chosen for two main reasons. The first reason was that the purpose was to conclude if paralinguistic cues can affect *accuracy in processing of spoken language*, not whether there are effects on processing that does not involve the linguistic stimuli (i.e., mere cued guessing). The second reason was the ambition to achieve high accuracy in the measurements (i.e., reliability).

When paralinguistic and prior context were manipulated, there was no contradictory information (i.e., either the extra information told the script, topic or emotional meaning, or there was no extra information). There were four reasons for this approach. Firstly, performance levels in visual and auditive speech processing tasks, such as those used in the studies of this thesis, are often close to floor effects even without false cues (e.g., Samuelsson, 1993). Secondly, to minimise the risk that participants would misconstrue the instructions as deceptive in conditions when they were not. Thirdly, to prevent that participants would use more guessing strategies than necessary. Finally, and essentially, not to risk lowering external validity as a consequence of the second and third argument.

## SUMMARY OF THE STUDIES

### Study I

*Purposes*

The purpose of Study I was threefold. Firstly, to examine if illustrating facially displayed emotions can facilitate visual speech processing, which was measured in three[16] levels. Secondly, to examine whether an effect of illustrating facially displayed emotions interacts with speech processing or skill levels. Thirdly, to examine how displayed emotions and speech processing levels with different task demands and different amounts and types of information to be processed affect confidence judgements.

*Method*

Twenty-seven normal-hearing adults were tested on four levels of topically cued visual speech processing in three levels of facially displayed emotion: positive, negative, and no (i.e., neutral) displayed emotion, in a balanced design. A human talker presented all stimuli, that were validated to be of either positive or negative valence in the scripts that were used. Two scripts were used in order to obtain valence effects of the words. Topical cueing was constant in order to facilitate speechreading, avoid floor effects, and increase external validity.

The speech processing levels were sentence-based speechreading (32 items), word decoding (32 items), word recognition (40 items), and word discrimination (40 items). The sentence-based speechreading and word decoding tasks used an open-set perceptual identification paradigm (Lively et al., 1994). The word recognition task used what can be denoted a visual–lexical matching paradigm, where the task was to decide whether a visually presented spoken target word matched a printed prime word, which could be identical, have a different beginning, or a different ending. The word discrimination task used a visual matching paradigm, where the task was to decide whether two visually and serially presented spoken words were identical or not, and where the pairs of words could be identical, differ in the beginning, or in the ending. All tasks had unique lists of stimuli. A secondary task was to rate confidence for each replication on a continuous scale ranging between one and seven.

Scoring in the sentence-based speechreading and word decoding tests was carried out by counting the number of correctly rendered word stems per subject and condition. In the word recognition and word discrimination tasks, the number of correct responses from the forced choices was counted. All data were then expressed as mean proportions correct per subject and condition.

*Results and discussion*

The original analyses of data showed no effect of facially displayed emotions on performance in any level of speech processing.

Interactions were found for skill level (cf. Just & Carpenter, 1992) and displayed emotion in word decoding and word discrimination. For word decoding, this suggests that speechreading skill is associated with ability to use paralinguistic information as constraints on activation spreading in the lexicon. With regard to word discrimination, where no lexical activation was required, but still may have been used, the result is more difficult to interpret. Lexical activation may have been used as a force of habit from previous perceptual identification tasks, especially as there were topical cues. If this was the case, the lexical constraint hypothesis may be true. If the participants did not try to identify the words, the results may suggest that articulatory distinctiveness may have been affected by the displayed emotions.

Confidence ratings were not affected by displayed emotion, but accuracy in the confidence ratings seemed to require lexical processing: confidence ratings and actual performance were only correlated in sentence-based speechreading and word decoding.

*Reanalysis of data*. The original analyses were performed with a separate one-way ANOVA per speech processing level. A drawback of this approach is that statistical power is reduced due to loss of reliability (i.e., more error variance), and that there is a greater risk of experimentwise Type I errors, as compared to performing an omnibus two-way ANOVA with speech processing level as a variable. The sentence-based speechreading test, the word decoding test, and the word discrimination test all measure aspects of visual speech processing. A reanalysis was therefore motivated. A 3 × 3 repeated measures ANOVA (Speech Processing Level × Displayed Emotion) was computed, yielding a main effect of displayed emotion, $F(2, 52) = 3.30$, $MSE = .014$, $p < .05$. Thus, presence of displayed emotions facilitated speech processing. A Tukey HSD test failed to yield significance for both valences, but a tendency toward significance for negative versus neutral valence was obtained, $p < .10$. This nominally larger improvement from displayed emotion for negative valence as compared to the nominal improvement for positive valence is interesting. Typicality ratings were lower for negative items, which suggests that they should be more difficult to identify due to weaker association with the script. However, cues from displayed emotion in combination with topical (script) primes may have resulted in higher cue distinctiveness (Hunt & Einstein, 1981; Hunt & Smith, 1996), thereby facilitating lexical access. There was also a main effect of speech processing level, $F(2, 52) = 389.73$, $MSE = .019$, $p < .001$, reflecting the fact that the word discrimination task is easier to perform than sentence-based speechreading and word decoding, with higher scores as a result.

*Additional data*. The word recognition test was included in the original design, but was not reported in the original article. The means for the word recognition task were $M = .91$ ($SD = .12$) for positive, $M = .86$ ($SD = .13$) for negative, and $M = .86$ ($SD = .09$) for neutral displayed emotion. When the word recognition task was included in the ANOVA, the main effects remained at the same significance levels. Displayed emotion yielded $F(2, 52) = 3.39$, $MSE = .016$, $p < .05$, and speech processing level $F(3, 78) = 517.23$, $MSE = .019$, $p < .001$.

When skill level (based on pooled speechreading performance, following the same procedure as in the article) was included in the ANOVA together with speech processing levels and displayed emotions, an interaction between skill level and speech processing levels appeared, $F(3, 72) = 4.57$, $MSE = .017$, $p < .01$. This reflects that visual speech processing skill above all is based on aspects of perceptual identification, where use of linguistic cues (cf. Boothroyd, 1988) as well as paralinguistic cues are suggested to be important. However, considering the means pattern, it appears that skill is also associated with lexical activation (which is required in all tasks except word discrimination), see Figure 2. Better speechreaders may be better at activating their visual and phonological representation of a target stimulus, thereby facilitating matching of stimulus to representation. In such a process, allocation of attention to key features of the representation and stimulus can be facilitated (cf. Samuel, 1990).

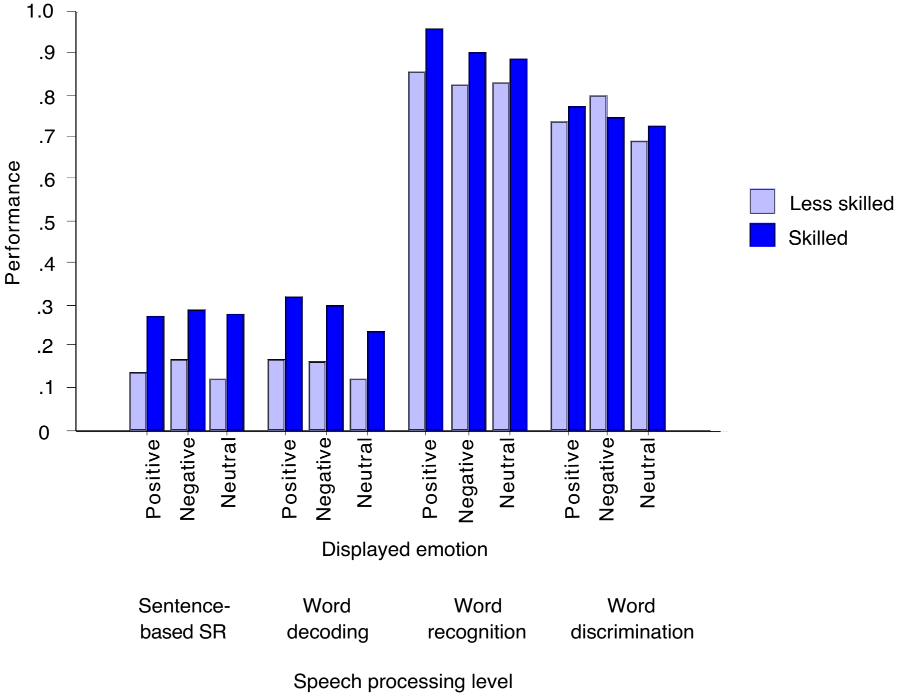

*Figure 2*. Performance as a function of speech processing and skill levels.

*Conclusions*. (1) Facially displayed emotions can facilitate speechreading, as demonstrated by the reanalysis. However, neither positive nor negative displayed emotion alone yielded significance as tested by Tukey HSD tests. (2) The illustrating facially displayed emotions appear to interact with topical cues in visual speech perception, thereby affecting cue distinctiveness, which may determine lexical access. (3) Skill in visual speech processing is suggested to be based on ability to integrate linguistic information, and paralinguistic information, such as illustrating facially displayed emotions. (4) Confidence ratings were unaffected by facially displayed emotions. (5) Accuracy in confidence ratings is suggested to require lexical activation.

## Study II

### *Purposes*

The first purpose of Study II was to examine whether facially displayed emotion can enhance performance in visual word decoding and sentence-based speechreading. (When Study II was conducted, the null hypothesis for displayed emotion in Study I was still considered to be intact.) The second purpose was to examine whether an effect of displayed emotion was mediated by linguistic complexity. The third purpose was to explore the association between verbal working memory span, verbal information processing speed, and speechreading performance.

### *Method*

Forty-eight normal-hearing adults were tested on topically cued visual sentence-based speechreading (144 items) and word decoding (152 items). A human talker presented all stimuli. There were two levels of facially displayed emotion (i.e., positive and negative displayed emotion vs. no displayed emotion), two levels of valence (i.e., positive and negative emotional meaning of the message), and two levels of linguistic complexity (i.e., long and short messages), in a balanced design. The effect of displayed emotions was thus more methodologically stringently tested as compared to Study I, where valence was not included in the analysis. The procedure for the speechreading tasks was essentially identical to that of Study I. The dependent measure was, however, the mean proportion of correctly rendered phonemes per participant and condition, in order to increase reliability compared to Study I.

The reading span test (Baddeley et al., 1985; Rönnberg et al., 1989) consisted of 54 three-word sentences, presented in groups increasing in span from three to six, with three sentences per span level. The task was to read aloud each sentence as it appeared in a word-by-word fashion on a computer screen, respond whether it made sense or not (half of the sentences did not), and finally to repeat either the first or last words of as many sentences as possible from the present trial in correct temporal order. The total number of correctly rendered words was measured.

The lexical decision-making test (Rönnberg et al., 1989) consisted of 100 three-letter words, 50 of which were common Swedish nouns, 25 were homophonous non-words, and 25 were non-homophonous non-words. The task was to indicate if the word was a real Swedish word or not by pressing keys as quickly as possible, but without making errors. Performance was measured as mean response time and number of errors.

*Results and discussion*

Facially displayed emotion improved speechreading performance for both sentences and separate words, in both conditions of linguistic complexity, and for both valences. Short stimuli yielded higher scores than long stimuli, and negative valence yielded higher scores than positive valence. There was an interaction between valence and displayed emotion, such that the largest facilitation of illustrating displayed emotion was obtained for utterances with negative valence. This pattern was also found in Study I. Also, as in Study I, the typicality ratings were lower for the items of negative valence. It can therefore be hypothesised that the relatively larger effect of negative displayed emotion is due to cue distinctiveness (cf. Hunt & Einstein, 1981; Hunt & Smith, 1996). Further, reading span but not lexical decision-making was significantly correlated with speechreading performance in all combinations of sentence-based speechreading, word decoding, long, and short messages.

*Conclusions.* (1) Visual speech processing benefits from paralinguistic information presented as illustrating facially displayed emotion. Both positive and negative displayed emotions can produce significant facilitation. (2) The magnitude of facilitation from displayed emotions on visual speechreading can be hypothesised as determined by cue distinctiveness, such that displayed emotions combined with topical cues constitute constraints on spreading activation. (3) Some aspect of working memory (e.g., verbal working memory span or attentional allocation) is associated with visual speech processing.

**Study III**

*Purposes*

The first purpose of Study III was to establish whether a synthetic talking head can be used for visual and audiovisual speechreading, especially if it can be used to enhance perception of poorly specified auditory speech. The second purpose was to explore whether topical cues (as prior context) produce different effects on visual and audiovisual speechreading as a function of the visual speech being presented by a human or by a synthetic talking head. Cue distinctiveness (cf. Hunt & Smith, 1996) was manipulated to this end.

*Method*

Ninety normal hearing adults were tested on a sentence identification test that consisted of 24 short sentences. Type of talker (i.e., human talker, or synthetic talking head, see Beskow, 1995, 1997) and modality (i.e., auditory, visual, or audiovisual presentation) constituted between-groups variables, resulting in five groups. The auditory speech was masked by broadband ”white” noise in the auditory and audiovisual modalities, with a signal-to-noise relationship of about –10 dB. Topical cueing (i.e., no, one, or two cue-words; cf. cue distinctiveness, Hunt & Smith, 1996) was within groups. The procedure was in many ways identical to that used in Studies I and II, with the exception of topical cueing being manipulated in three levels in Study III (i.e., no cue; topical cue; and topical cue plus cue to a content word, cf. Samuelsson & Rönnberg, 1993). The scoring procedure was the same as in Study II.

*Results and discussion*

The synthetic talking head was inferior to the human talker with regard to intelligibility overall, but the difference was emphasised in audiovisual speechreading. These results suggest that the visual speech of the synthetic talking head lacks some properties that are conveyed by human talkers, presumably due to programming of coarticulation. However, audiovisual presentation with both types of talker substantially raised performance compared to just listening to the speech in noise, indicating that both types of talker were successful in providing visually conveyed linguistic cues that complement the poorly specified auditory speech signal. Topical cue-word primes had the same type of facilitatory effect over all five blocks. This implies that topical cues are used in the same manner regardless of modality and type of talker. Thereby it is suggested that the effect of topical cues is to provide semantic constraints, and that the effect consequently is amodal and dissociated from the signal. The pattern of means suggests that cue distinctiveness (cf. Hunt & Smith, 1996) is strongly enhanced by topical plus specific cue-word primes, which effectively facilitates lexical access, replicating Samuelsson (1993) and Samuelsson and Rönnberg (1993).

*Conclusions*. (1) Synthetic talking heads can be used for conveying visual speech and for enhancing perception of poorly specified auditory speech. (2) Since the visual speech signal may be less specified than if conveyed by a human talker, additional information (e.g., from topical cues and linguistic cues from the auditory modality) may be required for obtaining intelligibility. (3) Topical cues are suggested to be used in an amodal code and utilised in the same fashion regardless of how the speech signal is presented. (4) Lexical access is suggested to be facilitated as a function of distinctiveness in topical cues in perception of poorly specified speech.

**Study IV**

*Purposes*

The purpose of Study IV was threefold. Firstly, to examine if displayed emotions affect impression of the talker and attitude to the task, which was visual speechreading. Secondly, to dissociate the effects from displayed emotion and linguistic specification on speechreading performance, attitude to the task, and impression formation. Thirdly, to test the hypothesis that displayed emotion affects speechreading accuracy via motivation, by exploring the associations between speechreading performance, impression of the talker, and attitude to the task.

*Method*

A total of 132 normal-hearing adults participated; 40 in Experiment 1, and 92 in Experiment 2. In both experiments, the task was topically cued sentence-based speechreading, and the procedure was essentially the same as in Studies I–III. The scoring procedure was identical to Studies II and III. After the speechreading test, impression of the talker was measured with a twenty-item questionnaire based on Asch's (1946) list of adjectives, and attitude to the task was measured with a seven-item questionnaire. A perceived extraversion index and an attitude index were formed from the questionnaire ratings.

*Experiment 1*. The speechreading test consisted of 20 sentences, spoken by a human talker. Displayed emotion (i.e., positive and negative displayed emotion vs. no displayed emotion) constituted between-groups variable, whereas valence (i.e., positive and negative emotional meaning of the message) was within groups.

*Experiment 2*. The speechreading test consisted of the 16 sentences from Experiment 1 that had produced the highest and lowest speechreading performance levels. Displayed emotion and linguistic specification (i.e., lower and higher) were between-groups variables, whereas valence was a within-groups variable.

*Results and discussion*

Study I (reanalysis) and Study II were replicated with regard to effect of displayed emotion, that is, displayed emotion again facilitated visual speech processing, and for both valences. Displayed emotion also affected attitude to the task (i.e., the attitude index) as well as impression of the talker (i.e., the perceived extraversion index). The participants were more positive toward the task and perceived the talker as more extraverted when he displayed emotions, compared to when he was constantly neutral. Linguistic specification also affected perceived extraversion, such that the talker was perceived to be less extraverted when he was less intelligible. No effect of linguistic specification was obtained for attitude to the task, indicating that attitude to the task is not associated with speech perception. The attitude index was associated with speechreading performance only when there was no displayed emotion and the stimuli were of lower specification. Perceived extraversion was not associated with speechreading performance in any instance. The combined results contradict the notion that displayed emotion affects speechreading accuracy by inducing emotion-related state in the perceiver, such as by increasing motivation. Therefore, it can be hypothesised that the facilitation from facially displayed emotions is due to either increased articulatory distinctiveness (i.e., higher specification of the visual speech signal) or to conveyance of semantic cues.

*Conclusions*. (1) Displayed emotions can positively affect attitude to the speechreading task. (2) Displayed emotions can affect person impressions, such that the talker is inferred to be more extraverted when emotions are displayed. (3) Linguistic specification can also affect person impressions: the talker was inferred to be more extraverted when the sentences were more intelligible. (4) Attitude to the task is generally not associated with speechreading performance. (5) Perceived extraversion is not associated with speechreading performance. (6) It is suggested that facially displayed emotions do not facilitate speechreading by affecting emotion-related states in the perceiver.

### Study V

*Purposes*

The first purpose of Study V was to test the hypotheses that displayed emotion facilitates perception of poorly specified speech (a) by conveying semantic cues and (b) by enhancing articulatory distinctiveness. The second purpose was to explore the effects of emotional and topical cues as functions of how the linguistic signal was presented. The third purpose was to examine the association between sensitivity to signal features (i.e., phoneme identification) and perception of poorly specified speech overall, and as an effect of stimulus specification and support for top-down influences.

*Method*

A total of 102 normal hearing adults participated; 30 in Experiment 1, and 72 in Experiment 2.

*Experiment 1*. A topically cued visual speechreading test which consisted of 36 sentences from one of the scenarios in Study I, presented by a human talker, was used. A balanced factorial repeated measures design comprised emotion (i.e., facially displayed emotion, emotional cue-words, and no emotional cues) and valence. The procedure was essentially the same as in Studies I, II, and IV. The scoring procedure was the same as in Studies II–IV.

*Experiment 2*. The same but four items from sentence-based speechreading and word decoding (both scenarios) in Study II were recorded with a new talker, resulting in a sentence identification test (72 items) and a word identification test (72 items). The procedure for presentation was in many respects the same as in Studies I–IV and Experiment 1, except for the fact that there were two independent variables that comprised cue-word primes. The scoring procedure was identical to Studies II–IV and Experiment 1. Modality (auditory, visual, or audiovisual presentation) constituted a between-groups variable. Linguistic complexity (sentences or separate words), type of talker (human talker or synthetic talking head), topical cue (absent or present), and emotional cue (none, facially displayed emotion, or emotional cue-word) constituted within-groups variables. With regard to the third purpose, modality and type of talker were included in the design to vary stimulus specification; linguistic complexity and topical cue to vary support for top-down influences.

The phoneme identification task consisted of 18 consonants presented in an /aCa/ format. There were 216 items in the visual and audiovisual presentations, and 108 items in the auditory presentation. Modality was a between-groups variable, whereas facially displayed emotion (positive, negative, and neutral) and talker were within-groups variables. Performance was measured as proportion of correct responses per condition.

*Results and discussion*

The semantic cue hypothesis was supported in both experiments: emotional cue-words increased visual speechreading performance. Due to the fact that the manipulation of displayed emotion was weak as revealed by a control of the stimulus material, there was no effect of displayed emotion – especially, the difference between neutral and negative facial display was not recognised in

Experiment 1. The specification of the displayed emotions as measured by ratings of magnitude and valence was reflected on speechreading scores. This finding closely resembles the effects of cue distinctiveness in Hunt and Smith (1996), and suggests that distinctiveness (or specification) in both topical cues (as prior context) and emotional cues (as paralinguistic context) determines lexical access.

The difference between the displayed emotion levels was, however, recognised in the synthetic talking head in Experiment 2. This was reflected in a tendency toward significance for the interaction between emotional cue and type of talker, where displayed emotion facilitated speechreading performance for the synthetic talking head. Experiment 2 revealed that emotional cue-word primes enhanced listening (auditory) performance as well as pooled visual and audiovisual speechreading performance, demonstrating that the effect is amodal.

Experiment 2 also showed that displayed emotions did not enhance phoneme identification for neither talker, but negative displayed emotion lowered phoneme identification for the human talker. This result, together with the fact that displayed emotion facilitated speechreading from the synthetic talking head, goes against the articulatory distinctiveness hypothesis. That is, speechreading performance can be enhanced by displayed emotions without increase in articulatory distinctiveness.

The human talker was superior only with regard to audiovisual phoneme identification, which indicates that some aspects that produce complementarity effects for natural audiovisual speech are lacking in the speech movements or synchronisation of the synthetic talking head. Word identification gave higher scores than sentence identification in all modalities. Topical cues facilitated speech perception in all modalities. The human talker gave higher speechreading scores overall, but especially in the audiovisual modality, as reflected by interactions between talker and modality.

There was an interaction between modality and linguistic complexity such that the difference between word identification and sentence identification was larger in the audiovisual condition. This may reflect a shift in response bias due to lexical and syntactic context (cf. Repp, Frost, & Zsiga, 1992; Samuel, 1990). An interaction between modality and topical cues suggested that topical cues have a larger facilitatory effect if more linguistic information can be perceived. Again, cue distinctiveness (cf. Hunt & Smith, 1996) can be hypothesised to be involved: the *general* topical cues can more effectively be used to access the lexicon if the *specific* cues (constituted by the perceived phonemes) are better specified, and vice versa.

An analysis of twin matched and pooled performance in the auditory and visual conditions versus performance in the audiovisual condition was made. Its results suggested that the linguistic constraints provided by sentence context and specification of the signal accomplished by the complementarity of (above all, natural) audiovisual speech interact. That is, topical cues, linguistic constraints, and phonemic specification may be used together, in a one-way activation (cf. Marslen-Wilson, 1990) or an interaction–activation (cf. McClelland & Elman, 1986) fashion. Once again specification of the signal appears to interact with cue distinctiveness (cf. Hunt & Smith, 1996), thereby influencing how efficiently the lexicon can be accessed. This time, specification of the signal was constituted by sentence versus word context, and bimodal versus unimodal presentation.

Phoneme identification and speechreading performance were correlated in both the visual and audiovisual modalities. Topical cues appeared to mediate the association between phoneme identification and speechreading performance for the synthetic talking head – the primes can be hypothesised to affect sensitivity to phonemic information. Part of this phenomenon may be attributed to perception of coarticulation: Samuel and Pitt (2003) found that compensation of coarticulation can be mediated by lexical activation.

*Conclusions.* (1) Displayed emotions can facilitate perception of poorly specified speech by conveying semantic cues. (2) The conveyance of semantic cues requires that the displayed emotions are sufficiently specified, such that they are identified. (3) Cue distinctiveness is suggested to be important for both prior context, such as from topical cues, and paralinguistic context, such as from facially displayed emotions. (4) Articulatory distinctiveness is suggested not to benefit from displayed emotions. (5) Speechreading can be facilitated by the talker's displayed emotions without increased articulatory distinctiveness as a byproduct of the displayed emotions. (6) Specification of the linguistic signal and constraints on activation spreading are suggested to affect how efficiently topical cues can facilitate access to the lexicon. (7) Phoneme identification is associated with perception of poorly specified speech. (8) Sensitivity to phonemic information is suggested to be influenced by priming.

## DISCUSSION

The results will be discussed under four main points, in accordance with the four purposes. The first three points are conclusions from the results, whereas the fourth is a proposed framework as a consequence of the conclusions.

## Conclusions

1. *Facially displayed emotions can facilitate visual and audiovisual speechreading*, as demonstrated in Studies I, II, IV, and V. Valence of the facially displayed emotion does not matter, that is, both positive and negative displayed emotions facilitate speech perception as long as the displayed emotions are recognised (Study V) and illustrate the emotional meaning of the utterance. The effect was found for four out of five talkers, one of whom being a synthetic talking head. Previous research had demonstrated that positive displayed emotion can facilitate visual speechreading (Johansson, 1997; Johansson & Rönnberg, 1996). The fact that negative displayed emotions can facilitate speechreading is a novel finding. Extra-facial gestures can also affect visual speechreading (Berger & Popelka, 1971; Popelka et al., 1971). Hence, two out of four forms of visual paralanguage in the taxonomy in Figure 1 have been empirically proven to affect visual speechreading. The remaining two instances (i.e., facial gestures and extra-facially displayed emotions) remain to be tested. As will be elaborated below, it is likely that facial gestures and extra-facially displayed emotions follow the same pattern as facially displayed emotions and extra-facial gestures – all forms of paralinguistic cues are suggested to facilitate perception of poorly specified speech, as long as they are in congruence with the linguistic signal, and are distinctive as cues.

2. *The facilitation of facially displayed emotions is due to conveyance of semantic cues*. There are five accounts for concluding that the displayed emotions facilitate perception by conveying semantic cues, whether or not improved articulatory distinctiveness and change in the state of the perceiver contribute. (However, no empirical indication of such contribution has been found.) Firstly, emotional cue-word primes and facially displayed emotions had the same type of facilitatory effect (Study V), even when articulatory movements were unaffected by the displayed emotions in a synthetic talking head (Study V). The facilitatory effect of emotional cue-words existed in the visual, audiovisual, and the auditory modalities (Study V). Both the synthetic talking head and the auditory modality eliminated the possibility that articulatory distinctiveness was affected. Secondly, phoneme identification was not enhanced by positive or negative displayed emotions in Study V. On the contrary, the human talker articulated less clearly when displaying negative emotions, but there was no improved specification from positive displayed emotion, as indicated by performance on the phoneme identification task (Study V). Thirdly, the associations between speechreading performance and attitude to the task, and between speechreading performance and impression formation, seem to be weak (Study IV; Lidestam et al., 2003). If emotional cues induce emotional states in

the perceiver and these states, in turn, affect speechreading performance, correlations with attitude to the task or person impression (such as perceived extraversion) are expected. Fourthly, the finding that negative as well as positive displayed emotion improve performance (Studies I, II, IV, & V) is contrary to the emotion-related state hypothesis – the valence of the displayed emotion does not matter as long as it is in consonance with the linguistic meaning of the utterance. However, although facially displayed emotions do not seem to affect speechreading accuracy via emotion-related states in the perceiver, the perceiver does seem to appreciate emotional expressiveness in the talker (Study IV; Lidestam et al., 2003). This is to say that this aspect of communication probably is important in real life communication – it usually feels more pleasant and stimulating to communicate with someone who expresses emotions than with someone who never does. Fifthly, and finally, the relatively larger enhancement from negative as compared to positive displayed emotion, in spite of lower typicality ratings for items of negative valence (Studies I & II), suggests facilitation of lexical access due to more distinct cueing (cf. Hunt & Smith, 1996). Cueing and lexical access imply semantic information; and as has been suggested, negative displayed emotions do usually not enhance articulatory distinctiveness. The combined data thus imply an effect of semantic priming. Note also that the data suggest that cue distinctiveness of topical cues interacts with specification of the phonemic information (Study V), linguistic context (Study V), and specification of paralinguistic cues (i.e., facially displayed emotion, Study V). It can therefore also be suggested that the phonemic information, the linguistic context, and paralinguistic cues *by themselves* constitute constraints on activation spreading, again implying semantic information (see also Boothroyd, 1988). Cue distinctiveness was explicitly tested in Study III.

Further support for the conclusion that facially displayed emotions facilitate speech perception by conveying semantic cues comes from the finding that extra-facial gestures can facilitate speechreading (Berger & Popelka, 1971; Popelka et al., 1971). This corroborates the notion that paralinguistic cues can enhance speechreading accuracy without improving articulatory distinctiveness and altering the emotion-related state in the perceiver, such as motivation. All of these empirical findings suggest that cues that can aid semantic (conceptual) classification, whether they be called pragmatic or paralinguistic, may facilitate lexical access.

The fact that the facilitation from paralanguage (i.e., facially displayed emotions and extra-facial gestures) is due to conveyance of semantic cues suggests that all forms of visual paralinguistic information can facilitate perception of poorly specified speech. Since the facilitatory effect of emotional paralinguistic information is due to the semantic properties of the signal, it is consequently also amodal. That is, we can expect facilitatory effects of paralanguage on language perception to exist regardless of how the language stimuli are conveyed, such as in auditory speech perception, reading, and sign language.

In accordance with this hypothesis, there are results from auditory speech perception and reading. In auditory speech perception, extra-facial gestures can affect responses to well specified speech, as was demonstrated in a Stroop-type paradigm by Langton, O'Malley, and Bruce (1996), suggesting that perceptual processing may be affected. There is further support from auditory speech perception, such that semantic transparency is a crucial factor in identification of morphologically complex words (Longtin, Segui, & Halle, 2003). Also, Wurm, Vakoch, Strasser, Calin-Jageman, and Ross (2001) showed that emotional paralinguistic information in the auditory mode facilitates auditory word recognition speed, and that this effect may be due to semantic priming, or expectation and amplification (Kitayama, 1990, 1991, 1996). In reading, valence judgements of words printed on pictures of faces are facilitated by congruent displayed emotions on faces used as background (Stenberg, Wiking, & Dahl, 1998). Interestingly, all of these studies of effects of paralinguistic cues on auditory speech perception used well specified, non-ambiguated speech stimuli; and the text stimuli in the task used by Stenberg et al. (1998) were also well specified. Thus, paralanguage can be hypothesised to affect language processing *in general* by providing semantic cues, such as by modifying the pragmatic aspects of the information. The effect is, however, especially pronounced for perception of poorly specified speech, where accuracy can be noticeably affected.

The effect of topical cues (Studies III[17] & V) has been demonstrated to be general. Studies III and V showed that the facilitatory effects followed the same pattern across all modalities, talkers (i.e., individuals) and types of talkers (i.e., humans and synthetic talking heads). Thus, we have converging operations (Garner et al., 1956) that increase the validity of the results. Other findings from speechreading (e.g., Lansing & Helgeson, 1995; Samuelsson, 1993), listening (e.g., Marslen-Wilson et al., 1996) and reading (e.g., Lucas, 1999; Meyer & Schvaneveldt, 1971; Neely, 1977, 1991; Tanenhaus & Lucas, 1987) corroborate the notion that the effect of topical cues as prior context is an effect on language processing *in general*. A final corollary conclusion is therefore that paralinguistic, linguistic, and prior context all serve the same purpose: to activate the lexicon and provide constraints on the spreading activation in the lexicon.

In sum, facially displayed emotions facilitate speechreading by conveying semantic cues, and are therefore semantic cues. It can be hypothesised that semantic cues in general (i.e., including paralinguistic cues, such as facially displayed emotions) can facilitate language perception in general; the effect is suggested to be universal across language processing. The mechanism for using all semantic cues is suggested to be constrained activation of the lexicon.

3. *Individual differences in information processing capacity are important with regard to perception of poorly specified speech.* As we have established, paralinguistic cues can facilitate speech perception, and especially perception of poorly specified speech. Skill level is associated with how well displayed emotion can be utilised in visual speech processing (Study I), which is a novel finding. Skill level is also associated with how well linguistic constraints can be utilised in speechreading (Study I; Boothroyd, 1988). Since paralinguistic and linguistic cues as well as topical cues all conclusively affect speech processing by conveying semantic information, it can be assumed that individual differences determine how all subcategories of these types of information are used separately as well as in combinations. Findings, such as, that differences between individuals with regard to working memory capacity are associated with written text processing (Just & Carpenter, 1992) suggest that the individual difference factor is inherent across all modalities and language processing tasks, and not only in visual speech processing.

Speechreading performance is correlated with working memory capacity as measured by the reading span test (Study II; Lyxell & Holmberg, 2000). However, conclusions about what aspects of working memory measured this way that are associated with speechreading are not possible to make on the basis of the reported data. As has been proposed (e.g., Lyxell & Rönnberg, 1993; Rönnberg, 2003*a*, 2003*b*; Rönnberg, Samuelsson, et al., 1998), it is possible that the association is due to perception of visual speech being so much based on top-down expertise in combining the multiple sources of information from prior and paralinguistic context, that the capacity to simultaneously store and manipulate information is taxed. Case studies (e.g., Lyxell, 1994; Rönnberg, 1993; Rönnberg et al., 1999) as well as correlations in group data (e.g., Study II; Lyxell & Holmberg, 2000) may be used as support for this explanation.

However, there are theoretical as well as empirical reasons to dispute the conclusion that successful speechreading requires a capacious working memory. First of all, memory span tasks undoubtedly tax the ability to simultaneously retain and manipulate information (i.e., the working memory span). Speechreading, on the other hand, whether it is of a short word or a sentence with ten words, semantically (conceptually) primed or not, cannot be claimed to tax working memory nearly as much (and certainly not the span). Priming of the lexicon is considered to be an automatic process, and, hence, does not strain working memory. The logical consequence is, therefore, that the constraints from top-down influences (e.g., priming) facilitate lexical identification, such that resources for postperceptual inferences are freed, not occupied. Further, completing typical speechreading items takes just a fraction of the time it takes to complete trials in the reading span task. This means that most (if not all) of the perceived phonological information can be retained in a phonological loop, relieving stress off a central executive (Baddeley, 2000; Baddeley & Hitch, 1974). Secondly, topically cued sentence-based speechreading, which has been suggested to tax working memory capacity, is not the only visual speech processing measure that is associated with measures of working memory. Decoding of short words is, too (Study II), and given the strong correlations between word decoding and sentence-based speechreading in Lyxell and Holmberg (2000), word decoding and reading span are likely to also be correlated in their study. Thirdly, the case study by Andersson and Lidestam (2003) demonstrated that a capacious working memory is not required for speechreading expertise: it can be based on superior bottom-up skills. In sum, working memory span does not appear to be a universally critical factor for successful speechreading.

An alternative interpretation is that attentional allocation (Samuel, 1990) is the key factor that explains the shared variance of the reading span task and speechreading tasks. Episodic encoding and retrieval are attention-demanding processes (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996) that are used in the reading span task. Attentional allocation can explain the association between memory span tests as well as how speechreading expertise can be based on superior bottom-up skills. The attentional allocation hypothesis can also explain correlations between speed in lexical access tasks and speechreading performance (Lyxell, 1994; Lyxell & Holmberg, 2000); facilitation of displayed emotion on visual word discrimination and word recognition; and the association between verbal cloze tests and speechreading. Samuel (1990) claimed that top-down information (i.e., sentence context) can facilitate attentional allocation such that perception is enhanced (i.e., that sensitivity to phonemes is improved). This claim is consistent with the results in this thesis: that the correlation between phoneme identification and sentence-based speechreading appears to be mediated by topical cues (Study V); and with the data pattern in Study II, where all levels of speech processing seem to be facilitated by facially displayed emotion. It also meshes well with the interaction of speechreading skill and displayed emotion in Study I. In fact, attentional allocation may even be a latent variable by increasing phonemic sensitivity in the association between phoneme identification and speechreading (Study V, Bernstein et al., 2000; Demorest et al., 1996). Further, it also adds to the explanation why readers with low reading span are unaffected in their parsing by the manipulation of sentence ambiguity (Just & Carpenter, 1992) – they fail to shift focus of attention. Such problems in ability to reallocate attention may well be an important factor in verbal working memory capacity, as measured by the reading span task. Boothroyd (1988) suggested that a larger difference in speechreading performance between highly competent and poor speechreaders of long as compared to short sentences (e.g., Hanin, as cited in Boothroyd, 1988) is likely to be due to differences in visual attention, short-term visual memory, and processing speed. The central role of attention in working memory was further emphasised by Engle (2002), and Kane and Engle (2002, 2003), who underlined that executive attention is crucial to working memory.

No data have been found on correlations between measures on working memory and lower levels of speech processing (e.g., word discrimination and word recognition). If the attentional allocation hypothesis is correct, correlations between tasks that can be hypothesised to measure attention (e.g., reading span performance) and lower levels of speech processing should exist. Stronger correlations between perception of more linguistically complex utterances and reading span could reflect that, for example, attentional allocation has a larger effect on processing of more complex messages (cf. Study II).

The data pattern of Study I, where all speech processing levels seemed to be facilitated by facially displayed emotion, suggests that the facilitation is not only due to postlexical inference processes (i.e., response bias), rather, there is a sensitivity effect. A facilitatory effect of displayed emotion in the word recognition and word discrimination tasks cannot be due to response bias, only to increased sensitivity. This result concurs with the mediating effect of topical cues on correlations between phoneme identification and speechreading in Study V, with the effect of sentence context on phoneme discrimination in Samuel (1981*a*), and with the word superiority effect in reading (Reicher, 1969).

To summarise, the combined empirical findings show that individual differences in information processing are associated with language processing in general, and processing of poorly specified speech in particular. Bottom-up as well as top-down processing capacity is associated with speechreading accuracy. Attentional allocation is suggested to account for these associations, by its suggested role in controlling the interaction between the linguistic signal and the top-down cues in the activation of the lexicon.

## A Conceptual Framework for Speech Perception

As has been established, paralinguistic context, prior context, and individual differences all affect perception of poorly specified speech. No theoretical models or frameworks that account for all of these effects have been found in the speech perception literature. Word recognition models (e.g., Forster, 1985; Gollan, Forster, & Frost, 1997; Klatt, 1979; Luce, 1987; Luce, Goldinger, Auer, & Vitevitch, 2000; Luce & Pisoni, 1998; Marslen-Wilson, 1984; Massaro, 1987, 1998; McClelland, 1991; McClelland & Elman, 1986; Morton, 1969, 1982; Norris, 1990) account for effects of linguistic cues, but do not explicitly incorporate the influence of neither semantic priming, paralinguistic cues, nor individual differences. One example is that top-down influence within the interactive-activation TRACE model (McClelland & Elman, 1986) is only triggered by the stimulus itself, allowing the identification of a stimulus word to affect sensitivity to phonemes backwards. Another example is that the neighborhood activation model (Luce & Pisoni, 1998) only includes word frequency as higher-level lexical information. Moreover, word recognition models appear to be based on the implicit assumption that the linguistic signal is well specified, such that all syllables, phonemes and phonemic features can be readily identified. Experimental stimuli have typically been well specified. Exceptions are, firstly, that in perceptual identification paradigms, masking has been used to degrade the signal in order to avoid ceiling effects (P. A. Luce, 1986, as cited in Lively et al., 1994). Secondly, in more widely used naming paradigms, the entire stimulus except one phoneme may be well specified (e.g., Dorman, Raphael, & Liberman, 1979; Marslen-Wilson et al., 1996; Massaro & Cohen, 1977; Samuel, 1981*a*, 1981*b*).

Word recognition models have, however, been fruitful in studies of intralinguistic phenomena, such as effects of neighbourhood density (e.g., Luce, 1987; Goldinger, Luce, & Pisoni, 1989; Luce & Pisoni, 1998), word frequency (e.g., Segui, Mehler, Frauenfelder, & Morton, 1982) phonological context (e.g., Massaro, 1989), lexical context (e.g., Elman & McClelland, 1988; Ganong, 1980), and form-based priming effects (e.g., Goldinger, Luce, Pisoni, & Marcario, 1992; Slowiaczek, McQueen, Soltano, & Lynch, 2000; Slowiaczek, Nusbaum, & Pisoni, 1987). Linguistic information can be hypothesised to be processed amodally, due to cross-modal and amodal effects (cf. Boothroyd, 1988; Lansing & Helgeson, 1995; Luce et al., 2000; Massaro, 1998). Consequently, these intralinguistic effects can be hypothesised to be inherent in processing of poorly specified speech as well. To conclude, word recognition models *may* explain intralinguistic effects in perception of poorly specified speech, but *do not readily* explain effects of semantic priming, paralinguistic cues, and individual differences.

Working memory models (cf. Baddeley, 2000; Baddeley & Hitch, 1974) serve the purpose of explaining how information is temporarily stored and manipulated. As has been argued above, the notion that working memory capacity (cf. Baddeley, 2000; Just & Carpenter, 1992) determines visual speechreading performance can be disputed. However, a central executive can be hypothesised to serve the function of utilising the constraints from semantic priming, linguistic, and paralinguistic context to make the most of the poorly specified linguistic signal properties. Attentional allocation (Samuel, 1990) can hypothetically be important in this complicated process.

To conclude: models of auditory speech perception are based on the assumption that the speech is *well* specified, can be considered to express domination of bottom-up processing, and typically only regard the speech stimuli as input. Further, because the stimuli are well specified, processing is highly automatised, and variation in individual differences is small. Processing of *poorly* specified speech, on the other hand, depends more upon top-down processing, the degree of automaticity may vary, as may individual differences. A comprehensive conceptual framework for speech perception, including perception of poorly specified speech, should therefore include the following 13 factors. (1) Prior context: constraints for the lexicon, preceding the signal. These constraints may, for example, be due to prior successful speech recognition, to topical or emotional priming from cue-words. (2) Distinctiveness in the cues from prior context. The distinctiveness determines how efficiently the lexicon is primed. (3) The core linguistic signal, that is, the morphology as it is presented via phonemes. (4) Specification of the core linguistic signal, that is, how accurately the phonemes and their features would be identified if all contextual cues were removed. (5) Intralinguistic constraints, for example, syntactic and phonological constraints, and lexical constraints, such as effects from neighbourhood density. (6) Paralinguistic constraints, such as emotional cues, gestures, and prosody. (7) Specification of the paralinguistic cues. For example, how well is the talker's emotion conveyed, and how explicitly are illustrators used? (8) Processing capacity for each stage. The capacity to process information can vary between individuals as well as between stages of processing within individuals. (9) Linguistic representations. The representations of words, grammar, phonology, et cetera, vary between individuals. These representations may also vary with regard to completeness within individuals. For example, phonological representations may be impaired, but semantic representations could be intact and diverse. (10) Paralinguistic representations. In the same way as linguistic representations, paralinguistic representations may vary between and within individuals. (11) Semantic representations. As concluded, both linguistic and paralinguistic constraints affect speech perception by conveyance of meaning. Therefore, semantic cues should not be regarded as an instance of linguistics nor paralinguistics, but as a separate and basic instance. Both linguistic and paralinguistic cues may be hypothesised to access the semantic representations as a result of activation spreading, and the semantic representations may therefore be hypothesised to have a crucial role in the lexicon. (12) Bimodality: natural audiovisual presentation results in a synergetic effect on perceived specification. (13) Strategies: automatic and conscious. The strategies can be assumed to determine what representations are activated and what properties of these representations that attention is allocated to.

Figure 3 is an attempt to present these considerations graphically. Note that this is a framework, not a model. This distinction is important, as models should provide testable predictions. No such claims can be made with a framework – its purpose is rather to provide a frame for building more specific models. Figure 3 is a schematic presentation of the information-processing of one individual. Variations in capacity between individuals are therefore not illustrated; variation within an individual could be expressed with different sizes on the components. Neither is flow of information within the working memory system and the long-term store indicated in the figure, for reasons of simplification.

The figure should be interpreted in the following way. First, there is prior contextual information, which precedes the signal. The distinctiveness of the prior contextual cues determines the priming effect on lexical access and identification. The core linguistic signal is embedded in a linguistic and paralinguistic context, and can be uni- or bimodally perceived. Accordingly, the signal may be uni- or bimodally evaluated and integrated (cf. Massaro, 1998). Amodal prior contextual information as well as the modal contextual information that embeds the core linguistic signal may affect attentional allocation and activate the lexicon (i.e., be instrumental in priming). According to this framework, the lexicon consists not only of linguistic information, but also of paralinguistic information. Further, semantic representations are regarded as the hub or the core of the lexicon, that is, as the basic property that integrates linguistic and paralinguistic information. When the signal enters the working memory system and consciousness, the information is handled by the executive function, where attention is situated. The executive function relies on slave systems to temporarily store information, and thereby frees resources for focusing attention to relevant properties of the signal. The executive function may also be hypothesised to rely on automatic (i.e., well practised) and conscious strategies for handling the combined information as efficiently as possible.

In sum, perception of poorly specified speech may be facilitated by a number of factors. The main conclusions were, firstly, that paralanguage in the form of emotional cues can enhance speech perception via facilitating lexical access to words (i.e., identification of a stimulus as a word that exists in the lexicon). Secondly, the facilitatory effect of emitted emotional cues is suggested to mediated by their degree of specification in transmission, their degree of ambiguity as percepts, and by how distinct the perceived emotions, coupled with topical cues, are as cues for lexical access. Thirdly, the facially displayed emotions affect speech perception by conveying semantic cues; no effect via enhanced articulatory distinctiveness, nor of emotion-related state in the perceiver is needed for facilitation. Fourthly, the combined findings suggest that emotional and topical cues can provide constraints for activation spreading in the lexicon. Finally, both bottom-up and top-down factors are associated with perception of poorly specified speech, indicating that variation in information-processing abilities is a crucial factor for perception if there is paucity in the sensory input.

To conclude, I suggest that within speech perception, the role of prior contextual cues and paralinguistic cues is to provide a *semantic frame* for the linguistic information, and that the semantic frame serves as constraints on the spreading activation in the lexicon. This simple conceptualisation can even be translated into old Gestaltist terminology: the joint information from the semantic frame and the linguistic signal is geared toward one Gestalt, based on both the linguistic cues, as figure, and the semantic frame, as background.

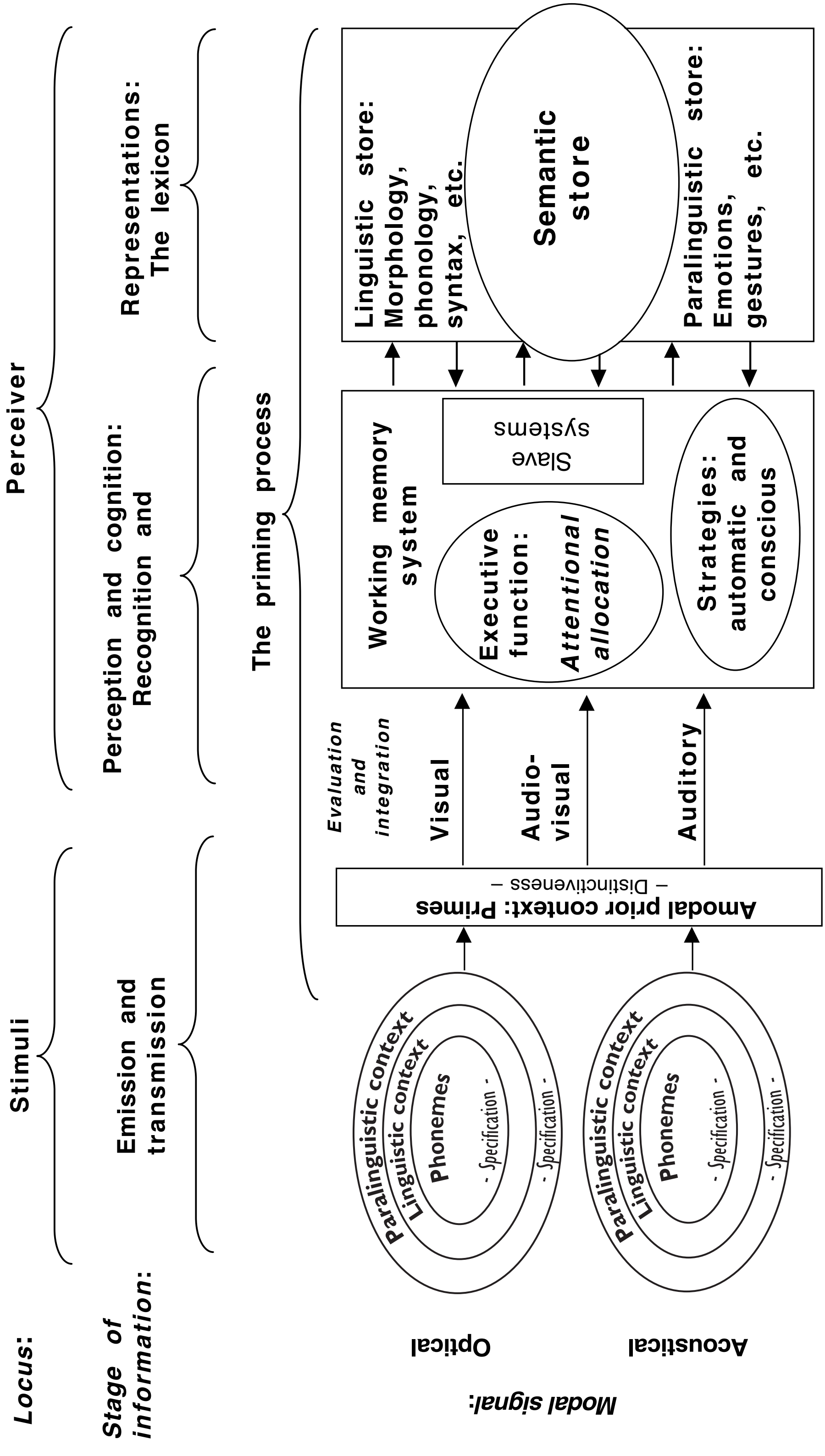
Locus:
Stimuli
Perceiver
Stage of information:
Emission and transmission
Perception and cognition: Recognition and
Representations: The lexicon
The priming process
Modal signal:
Optical
Acoustical
Paralinguistic context
Linguistic context
Phonemes
- Specification -
- Specification -
Amodal prior context: Primes
– Distinctiveness –
Evaluation and integration
Visual
Audio-visual
Auditory
Working memory system
Executive function: Attentional allocation
Slave systems
Strategies: automatic and conscious
Linguistic store: Morphology, phonology, syntax, etc.
Semantic store
Paralinguistic store: Emotions, gestures, etc.

*Figure 3*. A conceptual framework for speech perception.

### Directions for Future Research

The attentional allocation hypothesis needs to be tested. One consequence of efficient attentional allocation as a function of constraints on activation spreading is that sensitivity to the features of the signal is suggested to be enhanced (Samuel, 1990). Obviously, it would therefore be interesting to conduct an extended replication of Study I with controlled articulatory distinctiveness (e.g., by using a synthetic talking head) and manipulation of facially displayed emotion on a word recognition and a word decoding (identification) task. Response time can be included as a dependent variable, in order to obtain a more sensitive measure of lexical identification (cf. Lansing & Helgeson, 1995). Linguistic constraints and semantic priming may also be varied within the design, as may modality. Inclusion of verbal working memory tests such as the reading span test (Baddeley et al., 1985; Rönnberg et al., 1989) would allow testing of correlations with the speech processing tasks. The attentional allocation hypothesis predicts positive correlations between reading span and all speech processing tasks that require lexical activation (i.e., sentence-based speechreading, word decoding, and word recognition), but not necessarily with tasks that do not require lexical activation (i.e., word discrimination).

Paralinguistic cues and topical primes may be included in an experimental paradigm where words with ambiguous phonemes are to be identified, using signal detection theory to test effects of sensitivity and response bias (cf. Samuel, 1990). Visual or audiovisual presentation with attenuated auditory speech would allow natural ambiguity of phonemes. Increase in accuracy would be ascribed to increased sensitivity, whereas tendency to report words to be consistent with paralinguistic cues or primes would be regarded as reflecting response bias. Speeded-response lexical-decision paradigms may be used in the same design.

Facial gestures and extra-facially displayed emotions remain to be tested with regard to facilitation of perception of poorly specified speech. Further, the emotion-related state hypothesis can be explored with induced motivation (e.g., by rewards or by instructions) and induced affect (e.g., by music, cf. Niedenthal, et al., 1997).

The roles of vowels and syllables would be interesting to compare to the role of consonants in perception of visual speech (e.g., Study V; Bernstein et al., 2000). Visual identification of vowels is widely considered to be an easy task. However, considering the very poor specification of visual speech, it can be hypothesised that fast and accurate identification of the phonemes that are easiest to identify is crucial. Syllables have been suggested to be the key unit in speech perception (e.g., Mehler, Segui, & Frauenfelder, 1981; Segui, Dupoux, &

Mehler, 1990), and are also interesting due to the fact that syllables contain coarticulation. Coarticulation may be an essential factor in phoneme perception, as suggested by Studies III and V. Further, syllables often serve the function of morphemes. Therefore, by conveying morphological information, it can be hypothesised that syllables are crucial to perception of poorly specified speech due to their role in binding phonemic and morphological information and as entries to the lexicon (cf. Marslen-Wilson, Tyler, Waksler, & Older, 1994; Treiman & Danis, 1988). It has also been demonstrated that lexical activation (which may be affected by morphology) can mediate compensation for coarticulation (Samuel & Pitt, 2003). Inclusion of a syllable identification task in the aforementioned suggested designs may thus be one way of exploring how bottom-up and top-down information is integrated. One way of exploring where this integration is handled would be to incorporate neuroimaging techniques in the studies.

## REFERENCES

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist, 36*, 715–729.

Andersson, U. (2002). Deterioration of the phonological processing skills in adults with an acquired severe hearing loss. *European Journal of Cognitive Psychology, 14*, 335–352.

Andersson, U., & Lidestam, B. (2003). *Bottom-up driven speechreading in a speechreading expert: The case of AA*. Manuscript submitted for publication.

Andersson, U., & Lyxell, B. (1998). Phonological deterioration in adults with an acquired severe hearing impairment. *Scandinavian Audiology, 27*(Suppl. 49), 93–100.

Andersson, U., Lyxell. B., Rönnberg, J., & Spens, K.-E. (2001). Cognitive correlates of visual speech understanding in hearing-impaired individuals. *Journal of Deaf Studies and Deaf Education, 6*, 103–115.

Argyle, M. (1988). *Bodily communication* (2nd ed.). London: Methuen.

Arnold, M. B. (1960). *Emotion and personality: Vol. 1. Psychological aspects*. New York: Columbia University Press.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.

Auer, E. T. Jr., Bernstein, L. E., Waldstein, R. S., & Tucker, P. E. (1997). Effects of phonetic variation and the structure of the lexicon on the uniqueness of words. In C. Benoît & R. Campbell (Eds.), *Proceedings of the ESCA–ESCOP workshop on audio-visual speech processing* (pp. 21–24).

Auer, E. T., Bernstein, L. E., & Tucker, P. E. (2000). Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory & Cognition, 28*, 789–797.

Averill, J. R. (1980). A constructionist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience* (Vol. 1, pp. 305–339). New York: Academic Press.

Axtell, R. E. (1998). *The do's and taboos of body language around the world*. New York: Wiley.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). London: Academic Press.

Baddeley, A. D., Logie, R., Nimmo-Smith, I. & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24*, 119–131.

Berger, K. W. (1972). Visemes and homophenous words. *Teacher of the Deaf, 70,* 396–399.

Berger, K. W., & Popelka, G. R. (1971). Extra-facial gestures in relation to speechreading. *Journal of Communication Disorders, 3,* 302–308.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62,* 233–252.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America, 90,* 2971–2984.

Beskow, J. (1995). Rule-based visual speech synthesis. In J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, & F. J. Valverde (Eds.), *Proceedings of Eurospeech '95* (Vol. 1, pp. 299–302).

Beskow, J. (1997). Animation of talking agents. In C. Benoît & R. Campbell (Eds.), *Proceedings of the ESCA–ESCOP workshop on audio-visual speech processing* (pp. 149–152).

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). London: Academic Press.

Bode, D. L., Nerbonne, G. P., & Sahlstrom, L. J. (1970). Speechreading and the synthesis of distorted printed sentences. *Journal of Speech and Hearing Research, 13,* 115–121.

Boothroyd, A. (1988). Linguistic factors in speechreading. *The Volta Review, 90*(5), 77–87.

Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., et al. (2000). Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition and Emotion, 14,* 193–211.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77,* 305–327.

Cauldwell, R. T. (2000). Where did the anger go? The role of context in interpreting emotion in speech. In R. Cowie, E. Douglas-Cowie, & M. Schroder (Eds.), *Proceedings of the ISCA workshop on speech and emotion: A conceptual framework for research* (pp. 127–131). Belfast: Textflow.

Chomsky, N. (1957). *Syntactic structures*. The Hague, the Netherlands: Mouton.

Chomsky, N. (1980). *Rules and representations*. Oxford, England: Basil Blackwell.

Clouser, R. A. (1976). The effect of vowel–consonant ratio and sentence length on lipreading ability. *American Annals of the Deaf, 121,* 513–518.

Clouser, R. A. (1977, January). Relative phoneme visibility and lipreading performance. *The Volta Review, 79,* 27–34.

Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*, 113–139.

Conrad, R. (1977). Lipreading by deaf and hearing children. *British Journal of Educational Psychology*, *47*, 60–65.

Conrad, R. (1979). *The deaf schoolchild: Language and cognitive function*. London: Harper & Row.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In R. Cowie, E. Douglas-Cowie, & M. Schroder (Eds.), *Proceedings of the ISCA workshop on speech and emotion: A conceptual framework for research* (pp. 3–10). Belfast: Textflow.

Cowie, R. (2000). Describing the emotional states expressed in speech. In R. Cowie, E. Douglas-Cowie, & M. Schroder (Eds.), *Proceedings of the ISCA workshop on speech and emotion: A conceptual framework for research* (pp. 127–131). Belfast: Textflow.

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*, 159–180.

Crystal, D. (1997). *A dictionary of linguistics and phonetics*. Oxford: Blackwell.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, *19*, 141–177.

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. (1990). Approach–withdrawal and cerebral asymmetry: Emotional expression and brain physiology I. *Journal of Personality and Social Psychology*, *58*, 330–341.

Davidson, R. J., Mednick, D., Moss, E., Saron, C., & Schaffer, C. E. (1987). Ratings of emotion are influenced by the visual field to which stimuli are presented. *Brain and Cognition*, *6*, 403–411.

Demorest, M. E, & Bernstein, L. E. (1997). Relationships between subjective ratings and objective measures of performance in speechreading sentences. *Journal of Speech, Language, and Hearing Research*, *40*, 900–911.

Demorest, M. E., Bernstein, L. E., & DeHaven, G. P. (1996). Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with normal hearing. *Journal of Speech and Hearing Research*, *39*, 697–713.

Dimberg, U. (1997). Facial reactions: Rapidly evoked emotional responses. *Journal of Psychophysiology*, *11*, 115–123.

Dimberg, U., & Öhman, A. (1996). Behold the wrath: Psychophysiological responses to facial stimuli. *Motivation and Emotion*, *20*, 149–182.

Dorman M. F., Raphael L. J., & Liberman A. M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America, 65*, 1518–1532.

Ekman, P., & Friesen, W. V. (1972). Hand movements. *Journal of Communication, 22*, 353–374.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tartlatzis, I., Heider, et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotions. *Journal of Personality and Social Psychology, 53,* 712–717.

Elman, J., & McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23.

Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*, 413–422.

Erickson, & Schulkin, (2003). Facial expression of emotion: A cognitive neuroscience perspective. *Brain and Cognition, 52,* 52–60.

Esteves, F., Dimberg, U., & Öhman, A. (1994). Automatically elicited fear: Conditioned skin conductance responses to masked facial expressions. *Cognition and Emotion, 8,* 393–413.

Farah, M. J. (1989). Semantic and perceptual priming: How similar are the underlying mechanisms? *Journal of Experimental Psychology: Human Perception and Performance, 15,* 188–194.

Ferrand, L., & Grainger, J. (2003). Homophone interference effects in visual word recognition. *The Quarterly Journal of Experimental Psychology, 56A,* 403–419.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: The MIT Press.

Forster, K. I. (1985). Lexical acquisition and the modular lexicon. *Language and Cognitive Processes, 1,* 87–108.

Fridlund, A. J. (1991). Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology, 32,* 3–100.

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

Fridlund, A. J., & Gilbert, A. N. (1985, November 8). Emotions and facial expression. *Science, 230,* 607–608.

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

Ganong, W. F. III. (1980). Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review, 63*, 149–159.

Garstecki, D. C. (1976). Situational cues in visual speech perception by geriatric subjects. *Journal of the American Audiology Society, 2*, 99–106.

Garstecki, D. C., & O'Neill, J. J. (1980). Situational cue strategy influence on speechreading. *Scandinavian Audiology, 9*, 147–151.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove, England: Erlbaum.

Ghazanfar, A. A., & Logothetis, N. K. (2003, June 26). Facial expressions linked to monkey calls: Pulling a face to emphasize a spoken point is not seen as just a human prerogative. *Nature, 423*, 937–938.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*, 501–518.

Goldinger, S. D., Luce, P. A., Pisoni, D. B., & Marcario, J. K. (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1211–1238.

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in Hebrew-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1122–1139.

Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice* (2nd ed.). London: Routledge.

Harley, T. A. (1995). *The psychology of language: From data to theory*. Hove, England: Erlbaum.

Harré. R. (Ed.). (1986). *The social construction of emotions*. Oxford, England: Basil Blackwell.

Hergé (1962). *Tintin in Tibet*. London: Methuen.

Holmes, A., Vuilleumier, P., & Eimer, M. (2003). The processing of emotional facial expression is gated by spatial attention: Evidence from event-related brain potentials. *Cognitive Brain Research, 16*, 174–184.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*, 497–514.

Hunt, R. R., & Smith, R. E. (1996). Accessing the particular from the general: The power of distinctiveness in the context of organization. *Memory & Cognition, 24*, 217–225.

Hygge, S., Rönnberg, J., Larsby, B., & Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. *Journal of Speech and Hearing Research*, *35*, 208–215.

Isenberg, D., Walker, E. C. T., & Ryder, J. M. (1980, Fall). A top-down effect on the identification of function words [Abstract]. *Journal of the Acoustical Society of America*, *68*(Suppl. 1), S48.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*, 553–562.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception & Psychophysics*, *62*, 874–886.

Johansson, K. (1997). Speech gestures and facial expression in speechreading. *Scandinavian Journal of Psychology*, *37*, 132–139.

Johansson, K. (1998). A ″happy″ approach to speechreading (Doctoral dissertation, Linköping University, 1998). *Linköping Studies in Education and Psychology*, *57*.

Johansson, K., & Rönnberg, J. (1995). The role of emotionality and typicality in speechreading. S*candinavian Journal of Psychology*, *36*, 189–200.

Johansson, K., & Rönnberg, J. (1996). The role of facial approach signals in speechreading. *Scandinavian Journal of Psychology*, *38*, 335–341.

Johnson, H. G., Ekman, P., & Friesen, W. V. (1975). Communicative body movements: American emblems. *Semiotica*, *15*, 335–353.

Jones, S. E., & LeBaron, C. D. (2002). Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of Communication*, *52*, 499–521.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637–671.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Kitayama, S. (1990). Interaction between affect and cognition in word perception. *Journal of Personality and Social Psychology, 8*, 209–217.

Kitayama, S. (1991). Impairment of perception by positive and negative affect. *Cognition and Emotion, 5*, 255–274.

Kitayama, S. (1996). Remembrance of emotional speech: Improvement and impairment of incidental verbal memory by emotional voice. *Journal of Experimental Social Psychology, 32*, 289–308.

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics, 7*, 279–312.

Knapp, M. L., & Hall, J. A. (1997). *Nonverbal communication in human interaction*. Fort Worth, TX: Harcourt Brace Jovanovic.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt Brace.

Kuhl, P. K. (1991). Human adults and human infants show a ″perceptual magnet effect″ for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93–107.

Lamoré, P. J. J., Huiskamp, T. M. I., van Son, N. J. D. M. M., Bosman, A. J., & Smoorenburg, G. F. (1998). Auditory, visual and audiovisual perception of segmental speech features by severely hearing-impaired children. *Audiology, 37*, 396–419.

Langton, S. R. H., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1357–1375.

Lansing, C. R., & Helgeson, C. L. (1995). Priming the visual recognition of spoken words. *Journal of Speech and Hearing Research, 38*, 1377–1386.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research, 42*, 526–539.

Lemeignan, M., Aguilera-Torres, N., & Bloch, S. (1992). Emotional effector patterns: Recognition of expressions. *Cahiers de Psychologie Cognitive, 12*, 173–188.

Liberman, A. M. (1996). *Speech: A special code*. Cambridge, MA: MIT Press.

Lidestam, B., Beskow, J, & Lyxell, B. (2003). *Perceiving and relating to talking faces*. Manuscript in preparation.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 265–301). London: Academic Press.

Longtin, C.-M., Segui, J., & Halle, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes, 18*, 313–334.

Lucas, M. (1999). Context effects in lexical access: A meta-analysis. *Memory & Cognition, 27*, 385–398.

Luce P. A. (1987, Spring). The neighborhood activation model of auditory word recognition [Abstract]. *Journal of the Acoustical Society of America, 81*(Suppl. 1), S1–S2.

Luce, P. A., Goldinger, S. D., Auer, E. T. Jr., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62*, 615–625.

Luce, P. A., & Pisoni, D. A. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing, 19*, 1–36.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 122–147). Cambridge, MA: MIT Press.

Lundqvist, L.-O., & Dimberg, U. (1995). Facial expressions are contagious. *Journal of Psychophysiology, 9*, 203–211.

Lutman, M. E. (1983). The scientific basis for the assessment of hearing. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing science and hearing disorders* (pp. 81–129). London: Academic Press.

Lyxell, B. (1989). *Beyond lips: Components of speechreading skill*. Doctoral dissertation, University of Umeå, Sweden.

Lyxell, B. (1994). Skilled speechreading: A single case study. *Scandinavian Journal of Psychology, 35*, 212–219.

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11–14 years). *British Journal of Educational Psychology, 70*, 505–518.

Lyxell, B., & Rönnberg, J. (1987*a*). Guessing and speechreading. *British Journal of Audiology, 21*, 13–20.

Lyxell, B., & Rönnberg, J. (1987*b*). Necessary cognitive determinants of speechreading skill. in J. Kyle (Ed.), *Adjustment to acquired hearing loss* (pp. 46–54). Chippenham, England: Anthony Rowe.

Lyxell, B., & Rönnberg, J. (1989). Information-processing skill and speech-reading. *British Journal of Audiology, 23*, 339–347.

Lyxell, B., & Rönnberg, J. (1992). The relationship between verbal ability and sentence-based speechreading. *Scandinavian Audiology, 21*, 67–72.

Lyxell, B., & Rönnberg, J. (1993). The effects of background noise and working memory capacity on speechreading performance. *Scandinavian Audiology, 22*, 67–70.

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 148–172). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance: Vol. 10. Control of language processes* (pp. 125–150). London: Erlbaum.

Marslen-Wilson, W. D., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376–1392.

Marslen-Wilson, W. D., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review, 101*, 3–33.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989). Testing between the TRACE model and the Fuzzy Logical Model of Perception. *Cognitive Psychology, 21*, 389–421.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press, Bradford Books.

Massaro, D. W., & Cohen, M. M. (1977). The contribution voice-onset time and fundamental frequency as cues to the /zi/–/si/ distinction. *Perception & Psychophysics, 22*, 373–382.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1053–1064.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23*, 1–44.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McClelland, J. L., & Rumelhart, D. E., & The PDP Research Group (1986). *Parallel distributed processing: Vol. 2. Psychological and biological models*. Cambridge, MA: MIT Press.

McGurk, H., & MacDonald, J. (1976, December 23/30). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mehler, J., Segui, J., & Frauenfelder, U. (1981). The role of the syllable in language acquisition and perception. In T. F. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech*. Amsterdam: North-Holland.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*, 227–234.

Mogford, K. (1987). Lip-reading in the prelingually deaf. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 191–211). London: Erlbaum.

Montepare, J. M., Goldstein, S. B., & Clausen, A. (1987). The identification of emotions from gait information. *Journal of Nonverbal Behavior, 11*, 33–42.

Morton, J. (1969). Interaction of integration in word recognition. *Psychological Review, 76*, 165–178.

Morton, J. (1982). Disintegrating the lexicon: An information processing approach. In J. Mehler, E. Walker, & M. Garrett (Eds.), *On mental representation* (pp. 89–109). Hillsdale, NJ: Erlbaum.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226–254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336).

Niedenthal, P. M., Halberstadt, J. B., & Setterlund, M. B. (1997). Being happy and seeing "happy": Emotional state mediates visual word recognition. *Cognition and Emotion, 11*, 403–432.

Noble, W., & Perrett, S. (2002). Hearing speech against spatially separate competing speech versus competing noise. *Perception & Psychophysics, 64*, 1325–1336.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 87–104). Cambridge, MA: MIT Press.

Oatley, K. (1993). Social constructions in emotions. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 341–352). New York: Guilford.

Obusek, C. J., & Warren, R. M. (1973). Relation of the verbal transformation and the phonemic restoration effects. *Cognitive Psychology, 5*, 97–107.

Owens, E., & Blazek, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research, 28*, 381–393.

Pitt, M. A. (1995*a*). Data fitting and detection theory: Reply to Massaro and Oden. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1065–1067.

Pitt, M. A. (1995*b*). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1037–1052.

Platt, J. R. (1964, October 16). Strong inference. *Science, 146*, 347–353.

Popelka G. R., Lexington, K., & Berger, K. W. (1971). Gestures and visual speech reception. *American Annals of the Deaf, 116*, 434–436.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 274–280.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 243–335). New York: Academic Press.

Repp, B. H., Frost, R., & Zsiga, E. (1992). Lexical mediation between sight and sound in speechreading. *The Quarterly Journal of Experimental Psychology, 45A*, 1–20.

Rhodes, G., Parkin, A. J., & Tremewan, T. (1993). Semantic priming and sensitivity in lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 154–165.

Rhodes, G., & Tremewan, T. (1993). The Simon then Garfunkel effect: Semantic priming, sensitivity, and the modularity of face recognition. *Cognitive Psychology, 25*, 147–187.

Richardson, J. T. E., Engle, R. W., Hasher, L., Logie, R. H., Stoltzfus, E. R., & Zacks, R. T. (1996). *Working memory and human cognition*. New York: Oxford University Press.

Rönnberg, J. (1990). Cognitive and communicative function: The effects of chronological age and "handicap age". *European Journal of Cognitive Psychology, 2*, 253–273.

Rönnberg, J. (1993). Cognitive characteristics of skilled tactiling: The case of GS. *European Journal of Cognitive Psychology, 5*, 19–33.

Rönnberg, J. (2003*a*). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: A framework and a model. *International Journal of Audiology, 42*(Suppl. 1), 68–76.

Rönnberg, J. (2003*b*). Working memory, neuroscience, and language: Evidence from deaf and hard-of-hearing individuals. In M. Marschark & P. E. Spencer, *Deaf studies, language, and education*. Oxford, England: Oxford University Press.

Rönnberg, J., Andersson, J., Andersson, U., Johansson, K., Lyxell, B., & Samuelsson, S. (1998). Cognition as a bridge between signal and dialogue:

Communication in the hearing impaired and deaf. *Scandinavian Audiology, 27*(Suppl. 49), 101–108.

Rönnberg, J., Andersson, J., Samuelsson, S., Söderfeldt, B., Lyxell, B., & Risberg, J. (1999). A speechreading expert: The case of MM. *Journal of Speech, Language, and Hearing Research, 42*, 5–20.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual evoked potentials: Relation to adult speechreading and cognitive function. *Journal of Speech and Hearing Research, 32*, 725–735.

Rönnberg, J., Samuelsson, S., & Lyxell, B. (1998). Conceptual constraints in sentence-based lipreading in the hearing-impaired. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 143–153). Hove, England: Psychology Press.

Rönnberg, J., Öhngren, G., & Nilsson, L.-G. (1982). Hearing defiency, speechreading and memory functions. *Scandinavian Audiology, 11*, 261–268.

Rönnberg, J., Öhngren, G., & Nilsson, L.-G. (1983). Speechreading performance evaluated by means of TV and real-life presentation: A comparison between a normally hearing, moderately impaired, and profoundly hearing-impaired group. *Scandinavian Audiology, 12*, 71–77.

Russell, J. A., & Fehr, B. (1987). Relativity in the perception of emotion in facial expressions. *Journal of Experimental Psychology: General, 116*, 223–237.

Samuel, A. G. (1981*a*). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*, 474–494.

Samuel, A. G. (1981*b*). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1124–1131.

Samuel, A. G. (1990). Using perceptual-restoration effects to explore the architecture of perception. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 295–314). Cambridge, MA: MIT Press.

Samuel, A. G. (1991). A further examination of attentional effects in the phonemic restoration illusion. *The Quarterly Journal of Experimental Psychology, 43A*, 679–699.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*, 97–127.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48*, 416–434.

Samuelsson, S. (1993). Scripted knowledge packages: Implicit and explicit constraints on comprehension and memory (Doctoral dissertation, Linköping University, 1993). *Linköping Studies in Education and Psychology, 36.*

Samuelsson, S., & Rönnberg, J. (1991). Script activation in lipreading. *Scandinavian Journal of Psychology, 32,* 124–143.

Samuelsson, S., & Rönnberg, J. (1993). Implicit and explicit use of scripted constraints in lip-reading. *European Journal of Cognitive Psychology, 5,* 201–233.

Sanders, J. W., & Coscarelli, J. E. (1970). Relationship of visual synthesis skill to lipreading performance. *American Annals of the Deaf, 115,* 23–25.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.

Segui, J., Dupoux, E., & Mehler, J. (1990). The role of the syllable in speech segmentation, phoneme identification, and lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 263–280). Cambridge, MA: MIT Press.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia, 20,* 615–627.

Skinner, B. F. (1971). *Beyond freedom and dignity*. London: Jonathan Cape.

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G., & Lynch, M. (2000). Phonological representations in prelexical speech processing: Evidence from form-based priming. *Journal of Memory and Language, 43,* 530–560.

Slowiaczek, L. M., Nusbaum, H. C., & Pisoni, D. B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 64–75.

Smith, R. C., & Kitchen, D. W. (1972). Lipreading performance and contextual cues. *Journal of Communication Disorders, 5,* 86–90.

Sogon, S., & Masutani, M. (1989). Identification of emotion from body movements: A cross-cultural study of Americans and Japanese. *Psychological Reports, 65,* 35–46.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 11,* 361–407.

Stein, B. E., Wallace, M. T., Jiang, W., Jian, H., & Vaughn, W. (1999). Cross-modal integration: Bringing coherence to the sensory world. In D. W. Massaro (Ed.), *Proceedings of AVSP'99: International conference on auditory–visual speech processing* (pp. 23–28).

Stenberg, G., Wiking, S., & Dahl, M. (1998). Judging words at face value: Interference in a word processing task reveals automatic processing of affective facial expressions. *Cognition and Emotion, 12,* 755–782.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999, August 26). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400*, 869–873.

Summerfield, Q. (1983). Audio-visual speech perception, lipreading, and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing science and hearing disorders* (pp. 131–182). London: Academic Press.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). London: Erlbaum.

Tanenhaus, M. K., & Lucas, M. M. (1987). Context effects in lexical processing. *Cognition, 25*, 213–234.

Tillberg, I., Rönnberg, J., Svärd, I., & Ahlner, B. (1996). Audio-visual tests in a group of hearing-aid users: The effects of onset age, handicap age, and degree of hearing loss. *Scandinavian Audiology, 25*, 267–272.

Trager, G. L. (1958). Paralanguage: A first approximation. *Studies in Linguistics, 13*, 1–12.

Treiman, R. & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 145–152.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

van Son, N. J. D. M. M., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1993). Viseme classifications of Dutch consonants and vowels. *Journal of the Acoustical Society of America, 96*, 1341–1355.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20*, 130–145.

Wallbott, H. G. (1998). Bodily expression of emotion. *European Journal of Social Psychology, 28*, 879–896.

Walley, A. C., & Sloane, M. E. (2001). The perceptual magnet effect: A review of empirical findings and theoretical implications. In F. Columbus (Ed.), *Advances in psychology research: Vol. 4* (pp. 65–92). Hauppauge, NY: Nova Science.

Warren, R. M. (1970, January). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics, 9*, 358–362.

Warren, R. M., & Warren, R. P. (1970, December). Auditory illusions and confusions. *Scientific American, 223*(6), 30–36.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*, 672–683.

Williams, A. (1982). The relationship between two visual communication systems: Reading and lipreading. *Journal of Speech and Hearing Research*, *25*, 500–503.

Woodward, M. F., & Barber, C. G. (1960). Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, *3*, 212–222.

Wurm, L. H., Vakoch, D. A., Strasser, M. R., Calin-Jageman, R., & Ross, S. E. (2001). Speech perception and vocal expression of emotion. *Cognition and Emotion*, *15*, 831–852.

## NOTES

1. The *lexicon* can be conceived of as containing all the information regarding words, such as the *semantic*, *morphological*, *phonological*, *syntactic*, and *pragmatic* information (cf. Harley, 1995). Semantic refers to meaning; morphological refers to form and possible forms of words; phonological refers to speech sounds; syntactic to how words can be combined in sentences; and pragmatic to how the language is used.

2. As will be discussed, enhanced sensitivity to features as an effect of top-down influence of the stimuli could hypothetically be an alternative to inferences, especially if the signal is poorly specified (e.g., as in visual speech) rather than ambiguous (e.g., as in bad hand-writing).

3. The term *perception* has various meanings depending on research area. In this thesis, perception is used in the broad sense. It is acknowledged that top-down processes may influence perception in two ways: by enhancing sensitivity to the signal, as well as by aiding inferences at a later stage (that sometimes is referred to as *post-perceptual*, e.g., Harley, 1995). The dependent measure throughout this thesis was how accurately the phonemic information was rendered back. Speech perception may be contrasted with speech *understanding* (or speech *comprehension*), which entails the pragmatic aspect of how an utterance is interpreted in terms of, for example, the intention of the speaker.

4. The term *facially displayed emotions* is used here. In Studies II–V, the term was simply *displayed emotion*, whereas the term *emotional facial expression* was used in Study I. All three terms mean the same: *emotions that are facially displayed to illustrate the valence and emotional quality of the utterance*. The phenomenon under scrutiny was speech processing accuracy as an effect of facially displayed emotions, such that either facially displayed emotions illustrated the emotional meaning of utterances, or the same utterances were spoken with a neutral, nonillustrating facially displayed emotion.

5. The role of topical cues in all designs of the thesis was to provide a *prior context* (Harley, 1995) such that identification of the stimuli was facilitated. In this sense, the term *topical prime* would perhaps be more illustrative than topical cue, which does not indicate that the cue precedes the signal. However, the term topical cues will be used throughout the text for the sake of simplicity, and because the word *prime* is closely associated with traditional priming paradigms, where design, measurement, and analysis differ from those used in the present studies. Here, the term *topical* also refers to the cues to prior context in the studies, for sake of simplicity (see also Boothroyd, 1988). As variables, the

term *contextual cueing* was used in Study III, whereas *script* was used in Study V. Topical cues constituted a constant in Studies I, II, and IV.

6. *Linguistic context* comprises both *structural cues* (i.e., grammar, the inherent formal structure of language, which provides constraints for which combinations of components are possible or not, cf. Chomsky, 1957, 1980), and cues that activate stored representations of acquired knowledge of the language.

7. *Speechreading* performance refers to performance in the sentence-based speechreading task and the word decoding task (denoted the sentence identification task, and the word identification task, respectively, in Study V). That is, in order to denote the processing of visual or audiovisual speech speechreading, perceptual identification of the stimulus is required, and the stimulus has to have a meaning. Thus, the speech processing has to entail lexical identification. The task was open-set perceptual identification (cf. Lively et al., 1994) of either a sentence or a word at a time, and performance was measured as proportions of phonemes correctly rendered back.

8. It may be possible to test where and how integration of bottom-up and top-down information takes place in working memory models. However, this would probably require other experimental paradigms than those generally used in research on working memory to date (cf. Gathercole & Baddeley, 1993; Richardson et al., 1996). There appears to be a gap between research on working memory, on one hand, and research on access to lexical information, on the other hand. By closing this apparent gap, questions regarding integration of bottom-up and top-down information in relation to working memory may be answered.

9. *Perceptual distance* refers to the distance between the perceived phonetic information and stored prototypes (cf. Marslen-Wilson et al., 1996).

10. The term *synthetic talking head* is used here and in Study V, whereas *synthetic face* was used in Study III. However, the terms are equivalent as variables.

11. *Word decoding* refers to the auditory, visual, and audiovisual word identification tasks that were used.

12. *Valence* refers to the association of a concept along a positive–negative continuum (cf. pleasant–unpleasant, Argyle, 1988). Valence was used as a term in Studies IV and V, whereas *emotionality* and *emotional meaning* was used in Study I, and *hedonic impact* in Study II. As variables, the terms are equivalent.

13. The term *type of talker* is used here, and refers to the same variable as t*ype of face* in Study III, and *talker* in Study V. The comparison was between a human

talker and a synthetic talking head in both studies (although there were different talkers in Study III and Study V).

14. Here, the term *human talker* is used. It refers to the same variable as *natural face* in Study III, and *natural talker* in Study V.

15. In the article, the signal-to-noise relationship was reported to be approximately –4 dB. Subsequent behavioural data, however, suggest that this estimate was incorrect. The reported signal-to-noise relationship stems from the fact that there were two sources of noise, one of which embedded the speech, which made calculations difficult.

16. There were four levels of visual speech processing in the original design. However, only three of them were reported in the article, as the word recognition test was not reported.

17. Study III also explicated effects of cue distinctiveness in topical cues.