7

Summary and general discussion

The research described in the current thesis aimed to examine the ability of persons to comprehend degraded textual information and to evaluate its contribution to speech comprehension that is impeded by interfering noise and/or hearing loss. In the studies described in this thesis, we examined the speech comprehension benefit derived from partly incomplete visual information (masked text, Chapter 3) or partly erroneous text (text generated by means of Automatic Speech Recognition [ASR]). In the studies described in Chapters 4 and 5, we examined the objective benefit obtained from automatically generated captions ("subtitles") during speech comprehension in noise (Speech Reception Threshold [SRT] test). In Chapter 6, we evaluated the subjective benefit in more realistic listening tests. Figure 1 illustrates the main findings of the studies.

The results of the study described in **Chapter 2** showed that both the comprehension of masked text and the comprehension of speech in noise partly rely on modality-aspecific cognitive abilities. These abilities likely comprise inference-making skills that use the available linguistic information to complete partly incomprehensible sentences. In the study described in **Chapter 3**, we simultaneously presented masked text and speech in noise. The results showed that considerable speech comprehension benefit was obtained from relatively small amounts of visual information. The additional textual information provided extra linguistic context that enabled listeners to infer some of the missing words. The benefit obtained from partly masked written text in speech-in-noise comprehension was greater than the benefit predicted by the independent channels model described by Blamey, Cowan, Alcantara, Whitford, & Clark (1989); the mean difference between the observed and predicted audiovisual performances ranged between 15 and 25% correct.

The study described in **Chapter 4** focused on the benefit obtained from partly erroneous captions generated by ASR. The speech comprehension benefit was related to the ability to comprehend the captions when presented only visually. The higher readability of ASR *word*-output compared to the readability of ASR *phone*-output, and the generally higher word recognition performance indicated that ASR word-output is better suitable for supplementing speech comprehension than ASR phone-output. The benefit obtained from ASR output increased when the ASR accuracy was higher and when the delay of the text relative to the speech was shorter. For example, for ASR word accuracies around 75%, the benefit in SRT was approximately 3 dB SNR. The SRT benefit typically obtained from lipreading is around 5 dB SNR (Middelweerd & Plomp, 1987). Whereas an objective benefit in speech comprehension was obtained when captions were presented (Chapters 4 and 5), the findings in **Chapters 5 and 6** showed that the subjective effort and task load did not decrease when

the textual information was present. Hearing impaired listeners indicated that it was difficult to comprehend audiovisual information that is composed of degraded speech and erroneous and delayed text. In the study described in **Chapter 5**, the relatively high task load was reflected by higher effort ratings for the audiovisual tests than for the auditory tests, despite objectively measured speech comprehension benefit obtained from the visual information. This indicated that the perceived benefit obtained from the text may not outweigh the extra effort required to process the additional information. This finding underlines the need for applying both objective and subjective measures of speech comprehension benefit when evaluating assistive communication systems.

We did not observe a clear relation between age and the effort required to combine speech in noise and ASR-output. The results of the study presented in Chapter 5 suggested that working memory capacity and age are related to the subjective effort required to process the audiovisual information. A higher age was related to a lower working memory capacity and more effort required to combine the speech and text. The Text Reception Threshold (TRT) and the objectively measured speech comprehension benefit did not differ between the young and elderly listeners. The study described in Chapter 6 suggested that the reported task load decreases with increasing age. The difference in the findings of Chapters 5 and 6 could have been caused by differences between the subjective outcome measures and the different tasks performed by the participants. Furthermore, in Chapter 5, we calculated the relationship between age and the *difference* in the effort ratings for the auditory and audiovisual tests, whereas in Chapter 6, we used the 'absolute' task load ratings in the analyses. The working memory task applied in the studies in Chapter 5 and 6 (Spatial Span) was associated with the subjective effort ratings in Chapter 5. However, working memory capacity was not related to the objective speech comprehension benefit obtained from the text (Chapter 5) or the subjective task load ratings as described in Chapter 6. The working memory test mainly assessed the storage component of working memory. Various studies have shown that other, more complex working memory tasks that rely on both information processing and storage of the information in memory, like the Listening and Reading Span tests (Daneman & Carpenter, 1980; Rönnberg, 1990), are related to speech comprehension in various conditions (Pichora-Fuller, et al., 1995; Foo, Rudner, Rönnberg, & Lunner, 2007; Rönnberg, Rudner, Foo, & Lunner, 2008). Another explanation for the lack of association between the speech comprehension benefit and the working memory test used in this study may be the use of a *spatial* working memory test rather than a *verbal* working memory task. The relationship between the latter and (audiovisual) speech comprehension is

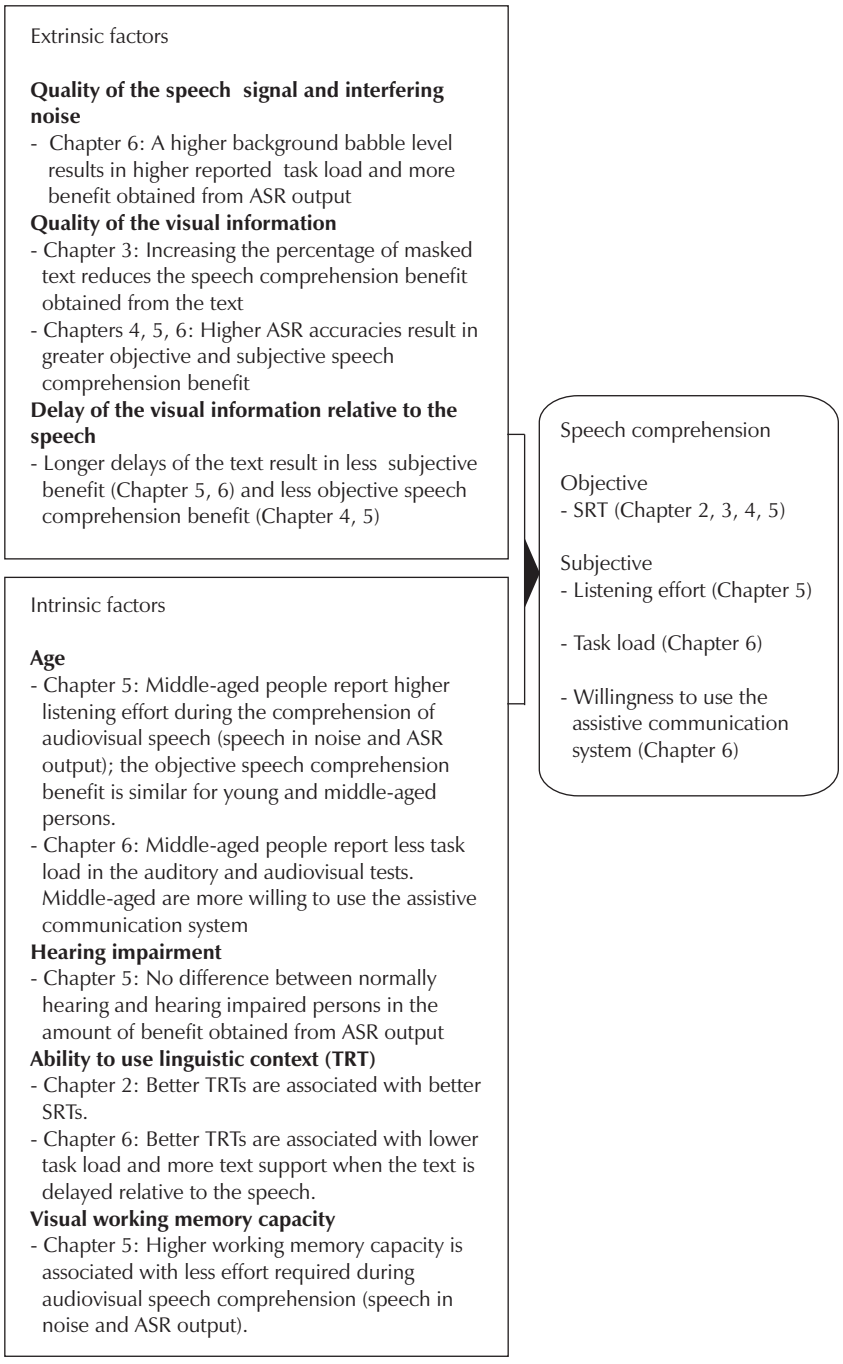

Figure 7.1. Overview of the main results of the studies described in the current thesis. ASR = Automatic Speech Recognition, TRT = Text Reception Threshold, SRT = Speech Reception Threshold.

likely more pronounced. Verbal working memory tests (relying on both the storage and the processing of perceptual information) are likely to explain more inter-individual variance in the ability to comprehend speech in noise with the support of erroneous textual information. One reason for applying the Spatial Span test in the current thesis was to obtain an estimation of working memory capacity not confounded by hearing loss (Van Boxtel et al., 2000; Valentijn et al., 2005). Secondly, we chose this specific working memory test as it does not rely on language comprehension processes similar to those needed to perform the TRT test. In other words, using different tests measuring different processes is more informative than applying two tests that to a large extent rely on similar abilities.

In **Chapter 6**, the hearing-impaired persons reported that it is effortful to simultaneously listen to speech in noise and to read partly incorrect text in order to fill in the incomprehensible parts of the speech. Longer speech fragments, low ASR accuracies (≤ 70%) and text delays resulted in subjective increments rather than decrements of the task load when additional visual information was presented. The fact that the comprehension of auditory and visual linguistic information partly depended on the same modality-aspecific processes may well have resulted in capacity shortage during the challenging task of comprehending telephone speech in background babble while reading partly erroneous and delayed text. Inference-making skills relevant for the completion of missing parts in linguistic information contribute to language comprehension in challenging conditions, but consume limited processing capacity. Thus, the reliance on such top-down processes is cognitively demanding and can limit speech comprehension if working memory capacity is overloaded by complex and effortful listening situations. If the subjectively perceived speech comprehension benefit obtained from the visual information does not outweigh the effort required to process the erroneous text, listeners may not be convinced that the text improves the comprehension of the speech. Reducing the task demands by reducing the number of errors in the text and by omitting the text delay frees up limited cognitive resources and thereby increases the subjectively indicated speech comprehension benefit derived from the text. However, the user evaluation performed in Chapter 6 indicated that high ASR performances do not guarantee a successful application of ASR in an assistive communication system.

Future studies to the application of ASR technology in assistive communication systems should focus on methods that reduce the complexity of the audiovisual task faced by the hearing impaired listener. Such methods contain adaptations of the ASR system that increase the recognition accuracy and reduce the

recognition delay. For example, instead of presenting the most probable recognition result, a list of the three most probable speech recognition results can be presented. This will increase the probability that the ASR output contains the correct transcription of the speech. However, presenting several ASR hypotheses for each spoken word or phrase will likely also increase the time required to read the ASR output and this may result in more effortful processing of the audiovisual information. Another adaptation of the ASR system that could improve the benefit obtained from the captions in speech comprehension is the use of *keyword spotting* instead of attempting to recognize each spoken word (e.g., Rose, 1995). In keyword spotting, the speaker is allowed unlimited use of words and phrasing, but the ASR system attempts to recognize only certain pre-defined (relevant) words from a limited vocabulary. In contrast to large-vocabulary ASR-systems, systems trained to only recognize certain keywords are able to reject out-of-vocabulary words, thereby reducing the ASR errors (El Méliani & O'Slaughnessy, 1997). Furthermore, it is important to minimize any differences between the data used to train the ASR system and the situation in which the ASR system is applied (Nusbaum, DeGroot, & Lee, 1995).

Optimizing the text display may also improve the readability of ASR output and the speech comprehension improvement obtained from the captions. In general, usability is increased when the user of the device can adjust the display settings. Like hearing loss, vision problems also increase with increasing age, thus text contrast and other aspects influencing text legibility are important to consider (Levitt, 1994). It should additionally be noted that even for high ASR accuracies, ASR output can still be quite difficult to read because of missing punctuation (capitals, periods, and commas). Inserting a line break each time the speaker pauses can increase the readability of the text (Bain, Basson, Faisman, & Kanevsky, 2005). Automatically highlighting particular words or phrases that were clearly emphasized in the speech may also facilitate the comprehension of the captions (Leitch and MacMillan, 2002). Although automatically inserting other text markers is difficult, any increase in the readability of the text by potential adaptations of the system and text display should be examined, as they may result in a less demanding audiovisual task.

Additionally, strategies can be developed to better prepare the users of the assistive communication system (i.e., the hearing impaired persons and the speakers whose speech is being recognized) to apply the system in daily conversations. Such strategies should improve the ASR accuracy and allow the hearing impaired listener more time to process the audiovisual information. The current thesis shows that these factors are crucial for the benefit obtained from the text. For example, the speaker whose speech is being recognized

by the ASR system could be trained to speak clearly and to pause between sentences (Shum, Myers, & Waibel, 2001). This will likely improve the ASR performance by facilitating the automatic detection of sentence endings and by making the speech less spontaneous. Providing feedback to the speaker may help them to optimize their speaking style for ASR (Kricos, 2006). This feedback information could contain the recognition output or a simplified indication of the ASR accuracy, like an indicator light that changes color if the recognition accuracy drops. Unfortunately, many aspects of human communication are so overlearned and automatic that they are difficult to modify (Karis & Dobroth, 1995) and processing the feedback information may slow down the conversation. Care has to be taken to provide reliable information on the ASR accuracy to the speaker, as unreliable speaker feedback (like unreliable confidence scores generated by the ASR system) may not improve the ASR accuracy and perhaps even distract the speaker.

In conclusion, reading incomplete text and speech comprehension in adverse listening conditions partly rely on the same, modality-aspecific cognitive abilities. Speech comprehension in noise by hearing impaired listeners improves by presenting masked or a partly erroneous transcription of the spoken sentences. However, if audiovisual speech comprehension is severely compromised (e.g., by hearing loss, background noise, and/or little time to process the audiovisual information), the task demands can exceed the available cognitive resources. In such cognitively taxing listening situations, the effort required to process the visual information may be high, thereby reducing the willingness of the listeners to use a system that provides imperfect textual information on the speech content. Although textual ASR output of near-future ASR systems likely will provide objective speech comprehension benefit during conversations, strategies should be developed to reduce the relatively high subjective work load associated with the processing of the textual information.