Research Article

Trends in Amplification
Volume 13 Number 1
March 2009 44-68

10.1177/1084713808330207
http://tia.sagepub.com
hosted at
http://online.sagepub.com

User Evaluation of a Communication System That Automatically Generates Captions to Improve Telephone Communication

Adriana A. Zekveld, PhD, Sophia E. Kramer, PhD, Judith M. Kessens, PhD, Marcel S. M. G. Vlaming, PhD, and Tammo Houtgast, PhD

This study examined the subjective benefit obtained from automatically generated captions during telephone-speech comprehension in the presence of babble noise. Short stories were presented by telephone either with or without captions that were generated offline by an automatic speech recognition (ASR) system. To simulate online ASR, the word accuracy (WA) level of the captions was 60% or 70% and the text was presented delayed to the speech. After each test, the hearing impaired participants (n = 20) completed the NASA-Task Load Index and several rating scales evaluating the support from the captions. Participants indicated that using the erroneous text in speech comprehension was difficult and the reported task load did not differ between the audio + text and audio-only conditions. In a follow-up experiment (n = 10), the perceived benefit of presenting captions increased with an increase of WA levels to 80% and 90%, and elimination of the text delay. However, in general, the task load did not decrease when captions were presented. These results suggest that the extra effort required to process the text could have been compensated for by less effort required to comprehend the speech. Future research should aim at reducing the complexity of the task to increase the willingness of hearing impaired persons to use an assistive communication system automatically providing captions. The current results underline the need for obtaining both objective and subjective measures of benefit when evaluating assistive communication systems.

Keywords: communication device for hearing impaired; assistive text display; automatic speech recognition; user evaluation

Persons with hearing impairments often experience difficulties in comprehending speech during telephone conversations. The difficulties can be attributed to several factors. Besides the reduced audibility of the speech caused by the hearing impairment, the limited bandwidth of the telephone signal (i.e., typically 300 to 3,400 Hz), and the fact that the face of the speaker is not visible, reduce telephone-speech comprehension (Kepler, Terry, & Sweetman, 1992; Milchard & Cullington, 2004). Kepler et al. (1992) performed a survey study in the United States to examine the problems with telephone conversations experienced by persons with hearing impairment. About 75% of the 104 respondents reported that speech comprehension was difficult, with a rating between "somewhat" and "extremely" difficult. About half of the participants used a hearing aid during telephone conversations, but 70% of them indicated that the coupling between

From the ENT/Audiology and EMGO Institute, VU University Medical Center, Amsterdam (AAZ, SEK, MSMGV, TH) and TNO Defence, Security and Safety, Soesterberg (JMK), The Netherlands.

Address correspondence to: Adriana A. Zekveld, ENT/Audiology and EMGO Institute, VU University Medical Center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands; e-mail: aa.zekveld@vumc.nl.

the hearing aid and the telephone was problematic, with the major problem being the "howling" when the telephone receiver was held close to the hearing aid. Additional reported sources of difficulty included problems adjusting the volume control and positioning the telephone properly to the ear without missing information. Less than 5% of the respondents indicated that they used the telecoil setting of the hearing aid to magnetically couple the hearing aid to the telephone. The respondents reported a strong need to improve speech comprehension during telephone conversations.

Assistive communication systems presenting visual information related to the speech content could improve speech comprehension, both for hearing aid users and for persons with speech comprehension problems who do not use hearing aids. An example of a system providing visual information for lipreading is Synface (e.g., Siciliano, Faulkner, & Williams, 2003). The Synface system uses automatic speech recognition (ASR) to recognize the spoken speech sounds. The recognized phones are used to control the lip movements of a synthetic face that is presented on a PC screen. An advantage of using ASR to recognize telephone speech is that special equipment is only needed on the side of the hearing impaired user. When manually corrected ASR output (i.e., an accurate phone-transcription of speech) was used to control the face expressions, the average speech comprehension benefit was about 22% for both normally hearing and hearing impaired listeners. Considerable interindividual variability in the comprehension benefit was observed though (Karlsson, Faulkner, & Salvi, 2003).

Another assistive technology displaying visual information on the content of the speech is the Liberated Learning Project (Leitch & MacMillan, 2003). This system has been specifically developed for use in lectures. It automatically generates a real-time speech-to-text transcription using ASR software trained on the lecturer's voice. Some students expressed concerns of erroneous information in their notes because of the ASR errors (Leitch & MacMillan, 2003). The students reported that the system improved teaching as long as the text was reasonably accurate (ASR accuracy >85%; Wald, 2006). Unfortunately, in most classroom environments, the accuracy level is <85% (Leitch, 2008; Leitch & MacMillan, 2003).

Compared with applications in which human operators transcribe the speech into text, advantages of using ASR technology in assistive communication devices are the potentially higher availability and the lower costs. Disadvantages of ASR technology are the missing punctuation in the text and the lack of indicators of speaker changes. Moreover, if ASR technology is used to extract the cues from the speech, the benefit obtained from the visual information will be limited by ASR errors (Levitt, 1994). Also, as the automatic processing of speech takes time, ASR inherently introduces a delay between the speech and the generation of the visual information. In earlier studies performed by our group, we examined the influences of ASR accuracy and text delay on the speech comprehension benefit obtained from automatically generated captions (i.e., textual transcriptions of speech obtained from an ASR system; Zekveld, Kramer, Kessens, Vlaming, & Houtgast, 2008, in press). Speech comprehension benefit was objectively measured by presenting auditory (audio only) and audiovisual (audio + text) speech reception threshold (SRT) tests (Plomp & Mimpen, 1979); the difference between the SRTs obtained in the audio-only and audio + text conditions was defined as the speech comprehension benefit. In addition, participants rated the effort required for combining the speech and the text (Zekveld et al., in press). Presenting the captions typically improved the SRT in noise by 1.5 to 2 dB speech-to-noise ratio (SNR), at the expense of more effort required in the audio + text conditions when the captions were delayed relative to the speech. The speech comprehension benefit was higher for better ASR accuracy levels and for shorter text delays.

Although the SRT tests used in our previous studies (Zekveld et al., 2008, in press) enabled the systematic manipulation of the ASR accuracy and the text delay, they did not realistically simulate the application of ASR during telephone conversations. In the SRT tests, short sentences were presented in isolation. If participants have to comprehend longer speech segments (several utterances) while concurrently reading the corresponding captions that are delayed relative to the speech, it may be relatively difficult to combine the speech and the text. The benefit obtained from captions generated by means of ASR therefore needs to be examined in more realistic tests in which longer speech utterances are presented.

In the Zekveld et al. (in press) study, we mainly focused on objective measures of speech comprehension benefit obtained from the captions. Next to objective measures, subjective measures are needed, as people will not use new technologies if they do

not believe they benefit from it, despite objectively measured speech comprehension benefit (Boothroyd, 2004; Jerger, Chmiel, Florin, Pirozzolo, & Wilson, 1996; Kricos, 2006). Thus, even if a system significantly improves speech comprehension, the subjectively perceived benefit may be low, which will reduce the willingness to use the system (Nusbaum, DeGroot, & Lee, 1995). User trials with a prototype or a simulation of the assistive communication system are therefore indispensable (Karis & Dobroth, 1995). The current study examined the subjective speech comprehension benefit from automatically generated captions and evaluated the willingness of persons with hearing loss to use future automatic captioning technology.

In a pilot study, we examined the performance of several Dutch ASR systems (research and commercial systems) for the recognition of spontaneous telephone conversations. For natural telephone conversations, the performance of the ASR systems was highly variable: Some speakers and utterances were recognized well, whereas other parts of the conversations resulted in many recognition errors. Furthermore, the overall performance appeared to be insufficient for our research goals. For this reason, we did not test the benefit obtained from textual ASR output during spontaneous bidirectional telephone conversations. Instead, participants listened to short stories presented by telephone while reading captions that were generated offline by an ASR system. Only stories that yielded sufficient ASR performance (i.e., ASR accuracy >60%) were presented. Either speech was presented alone (audio only), or both speech and text were presented (audio + text). The subjective benefit, or task load decrement, obtained from the captions was measured by means of the widely used NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988), a reliable, multidimensional task load rating scale (Hill et al., 1992). The NASA-TLX has been widely used for measuring the workload imposed by using ASR technology in military vehicles (Leggatt & Noyes, 2004) and for evaluating bimodal interfaces (Murata, 1999). The participants of the current study also evaluated the support obtained from the captions and the problems with the text delay and ASR errors. To mimic realistic situations and to prevent a too easy listening task, interfering speech (babble noise) presented in the test room masked the target speech presented by telephone. The ability to obtain benefit from the partly erroneous text may be related to cognitive functions like working memory and linguistic abilities. In the group of normally hearing participants included in our previous study (Zekveld et al., in press), a higher age and lower spatial working memory capacity were associated with more effort required for combining the text and the speech. Unfortunately, the hearing impaired participants in that study did not perform the working memory task, thus for this group, the relation between working memory capacity and the ability to benefit from the text remained unclear. In the current study, we therefore included the working memory test and a test of the ability to complete partly, visual masked, written sentences and examined the relation between age, the performance on these tests and the subjectively evaluated benefit from the captions.

Methods Experiment 1

Participants

Participants were patients of the audiological center of the VU University Center with hearing impairments. They had consulted the audiology department for a hearing aid prescription or a regular hearing assessment. Those who reported speech comprehension problems during daily telephone conversations were asked to volunteer. Other inclusion criteria were that patients used spoken language to communicate in daily life, that they did not use sign language, and were aged between 18 and 85 years. Cochlear implant users were not included, as they would form a rather distinct and small subgroup. We applied no other inclusion criteria regarding the degree and type of hearing loss or hearing aid use. A total of 20 patients (9 female, 11 male) were tested in Experiment 1. Their ages ranged from 32 to 82 years, with a mean age of 60 years (SD = 10.8 years). Pure-tone thresholds and speech discrimination data were available for all patients. Speech discrimination was measured unaided by presenting lists of 10 Dutch CVC (consonant–vowel–consonant) words monaurally at a number of sound levels. The percentages of identified phonemes were scored separately for both unaided ears. The maximum word discrimination of the best ear had to be at least 80% for inclusion in the current study, to ensure that speech comprehension was sufficient to perform the tests. Most participants had sensorineural hearing loss, but five participants had mixed sensorineural/conductive hearing loss.

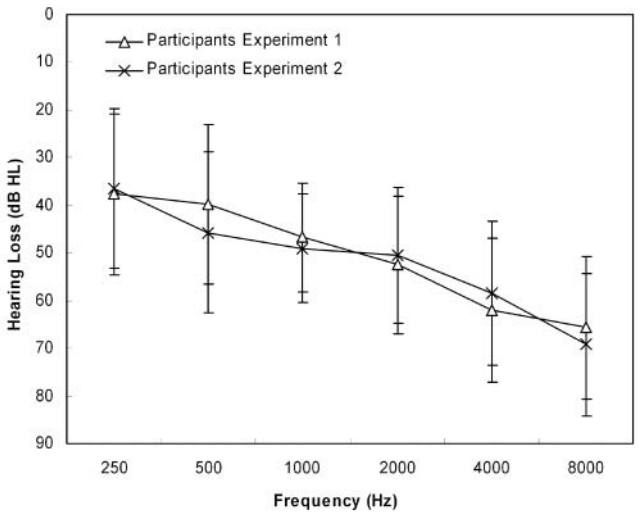

Figure 1. Means and standard deviations (error bars) of the unaided pure-tone audiometric thresholds (averaged over both ears) of the participants.

Figure 1 shows means and standard deviations of the pure tone hearing thresholds (average of both ears). In all, 13 participants used two hearing aids, and 3 participants used one hearing aid in daily face-to-face listening situations. Six participants used a hearing aid during telephone conversations; one of them used a special telephone program of the hearing aid. None of the participants used the telecoil setting of their hearing aid. The remaining participants did not use a hearing aid, either because they used their better, unaided ear during telephone conversations or because of interference from feedback sounds (howling) caused by the telephone receiver. Four participants used special telephone equipment such as telephone amplifiers. All participants were native Dutch speakers who reported normal or corrected-to-normal vision, no dyslexia, and no history of neurological disease. All provided written informed consent in accordance with the Ethical Committee of the VU University Medical Center.

Stimuli and Tests

Table 1 presents an overview of the tests presented in the current study. The Telephone-Narrative Comprehension in Babble Noise (Telephone-NC in babble noise) test was the focus of the current study; it examined the speech comprehension benefit obtained from the captions. In this test, auditory narratives were presented by telephone, and interfering babble noise was presented in the test room. Two initial tests were performed to set the appropriate babble-noise levels in the Telephone-NC in babble-noise test. The Telephone-Speech Reception Threshold Test in Quiet (Telephone-SRT in quiet) was used to determine the appropriate presentation level of the auditory stimuli. The Telephone-Speech Reception Threshold Test in Babble Noise (Telephone-SRT in babble noise) estimated the individuals' ability to comprehend speech by telephone in babble noise (the Telephone-Speech to Babble-Noise Ratio [SNR] corresponding to 50% sentence intelligibility). In each listening test, the target speech was presented by telephone. In the Telephone-SRT and Telephone-NC in babble-noise tests, interfering babble noise was presented in the test room. Additional tests that were performed included the Text Reception Threshold (TRT) test (Zekveld, George, Kramer, Goverts, & Houtgast, 2007), and the Spatial Span (SSP) visual working-memory test (Cambridge Neuropsychological Test Automated Battery [CANTAB]; Owen, Downes, Sahakian, Polkey, & Robbins, 1990). The TRT test is a visual analogue of the SRT test in noise, and measures abilities involved in the completion of partly, visually masked, written sentences that are relevant for the comprehension of speech in noise (George et al., 2007; Zekveld et al., 2007). We presented the TRT test to examine whether the linguistic abilities involved in comprehending incomplete textual information, as measured by the TRT tests, are also relevant for the comprehension of partly incorrect captions.

Besides the amount of hearing loss of the measured ear, several other factors can substantially influence telephone-speech comprehension when babble noise is presented in the room. First, the positioning of the telephone receiver (e.g., the pressure of the telephone receiver against the ear) influences the audibility of both the telephone speech and the babble noise. Second, the use of a hearing aid and the positioning of the telephone receiver relative to the hearing aid also influence speech comprehension and the audibility of the babble noise. Third, the severity of the hearing loss of the ear not used to listen to the telephone-speech influences the audibility of the babble noise: More severe hearing loss will result in less interference from the babble noise. Hence, interindividual differences in the ability to comprehend the telephone speech depended

Table 1. Tests Presented in the Current Study

	Abbreviation	Aim	Outcome Measure
Preparatory tests			
1. Telephone-speech reception threshold in quiet	Telephone-SRT in quiet	Determination of the appropriate babble-noise level in the Telephone-NC_{BB} tests	Telephone-SRT in quiet (dB SPL)
2. Telephone-speech reception threshold in babble noise	Telephone-SRT in babble noise	Estimation of the comprehension of telephone speech in babble noise; used for adapting the babble-noise level in the Telephone-NC in babble-noise test	Telephone-SRT_{BB} (dB SNR)
Main tests			
3. Telephone-narrative comprehension in babble noise: audio only and audio + text	Telephone-NC in babble noise	Examination of the speech comprehension benefit obtained from automatically generated captions during telephone speech comprehension in babble noise. Word accuracy 60% and 70%	NASA-TLX and ASR-output evaluation (rating scales)
4. Interval telephone-narrative comprehension in babble noise	Interval Telephone-NC in babble noise	Examination of the speech comprehension benefit obtained from automatically generated captions during telephone speech comprehension in babble noise. Word accuracy 44% to 88%	NASA-TLX performance subscale and text support rating
Cognitive/linguistic tests and questionnaire			
5. Text reception threshold	TRT	Examination of the ability to complete masked sentences	TRT (% unmasked text)
6. Spatial span	SSP	Examination of visual working memory capacity	Working memory span
7. Questionnaire		Part 1: Evaluation of daily telephone speech comprehension problems Part 2: Evaluation of captions presented in the tests	

NOTES: SNR = speech-to-noise ratio; NASA-TLX = NASA Task Load Index.

on a number of factors besides the intelligibility of the speech. Therefore, the Telephone-SRT in quiet and Telephone-SRT in babble noise thresholds were used as rough estimations of the intelligibility of the speech in the Telephone-NC in babble-noise tests, as they do not reflect "pure" telephone-speech comprehension ability. Participants were instructed to simulate their daily telephone conversations regarding the ear used to listen to the telephone speech, the use of hearing aid(s), and the positioning of the telephone receiver relative to the ear/hearing aid.

Apparatus

A Dell OptiPlex 6X745 Intel core 2 DUO E6700 desktop PC with a touch-sensitive ELO-C1728 CRT monitor controlled stimulus presentation during all tests. Sound files were stored on a computer hard disk. The speech was band-pass filtered from 300 to 3,400 Hz and A-law companded by computer software to simulate the public switched telephone network. Speech signals were generated by a Creative Audigy SoundBlaster 2 ZS sound card. The output of the sound card was directly connected to the speaker of the telephone (Krone desk phone, type T65 TDK). Thus, typical public switched telephone network characteristics were "simulated" by software processing of the signals. Sound calibrations of the telephone speech were performed with a Brüel & Kjær 2260 Observer, a Brüel & Kjær Artificial Ear (type 4152), a flat-plate adaptor for this artificial ear, and a soft seal to prevent sound leakage between the

flat-plate adaptor and the telephone handset. The babble noise was generated with a SoundMAX integrated digital audiocard and two Bowers & Wilkins 200 series speakers coupled to a Bryston Ltd. 2B power amplifier. Sound calibrations of the babble noise were performed with a Brüel & Kjær 2260 Observer.

Telephone-Speech Reception Threshold in Quiet

The Telephone-SRT in quiet test adaptively estimated the sound level of telephone speech (dB SPL) required for reproducing 50% of the sentences without error. One list of 13 short, everyday Dutch sentences (Plomp & Mimpen, 1979) was presented. Participants were asked to repeat each sentence and encouraged to make their best guess for sentences for which they were not sure. A sentence was scored as correct if the participant was able to repeat each word of the sentence without error. No feedback was given during the tests. The first sentence was presented at a sound level below threshold and was presented repeatedly, increasing the sound level (in increments of 4 dB SPL), until the participant repeated the sentence correctly. Each of the following sentences was presented once; if the sentence was identified correctly, the sound level of the next sentence was decreased by 2 dB SPL; if it was identified incorrectly, the sound level was increased by 2 dB SPL. The Telephone-SRT in quiet was the mean sound level of sentences 5 to 14. The purpose of estimating the Telephone-SRT in quiet was to determine the appropriate babble-noise levels for the Telephone-SRT and Telephone-NC in babble-noise tests (see Procedure section).

Telephone-Speech Reception Threshold in Babble Noise

The Telephone-SRT in babble-noise test used the same adaptive procedure as applied in the Telephone-SRT in quiet test (Plomp & Mimpen, 1979), but now babble noise presented in the test room served as masker (for levels: see Babble Noise section). Thirteen short, everyday Dutch sentences (Plomp & Mimpen, 1979) were presented in each test. We estimated the speech to babble noise ratio (SNR in dB) at which participants reproduced 50% of the sentences without error. The SNR was varied adaptively by changing the telephone-speech level. The Telephone-SRT in babble noise was the mean SNR of sentences 5 to 14. Lower thresholds in babble noise indicate better performance. The Telephone-SRT in babble noise was used to estimate the individuals' ability to comprehend telephone speech in babble noise. Based on the Telephone-SRT in babble noise, for several participants, the level of the telephone speech in the Telephone-NC in babble-noise tests was adapted (see Procedure section).

Babble noise. The babble noise presented during the Telephone-SRT and Telephone-NC in babble noise tests consisted of four-speaker babble. Versfeld, Daalder, Festen, and Houtgast (2000) developed the speech material. About 1,000 sentences were available for each of the four speakers (two males, two females). The speech material of the speakers was mixed and presented without pauses between the sentences. The audio track was looped at the end of the audio file. Two loudspeakers presented the babble noise; the audio track played by one of the loudspeakers was delayed by several seconds compared with the track played by the other one, to simulate babble noise containing eight voices. The two loudspeakers were located at a distance of approximately 1.5 metres in the front-left and front-right (45°) of the participant. The babble-noise level was either 60 or 70 dB SPL (depending on the individual Telephone-SRT in quiet), and will be referred to as the high babble-noise level. The low babble-noise level was 6 dB below the high babble-noise level (i.e., 54 or 64 dB SPL, see Procedure section). None of the sentences in the babble noise was used in the Telephone-SRT tests.

Telephone-Narrative Comprehension in Babble Noise

Stimuli: narratives. The Dutch Office of Intercultural Evaluation (Bureau InterCulturele Evaluatie, ICE) developed the auditory narratives. Dutch educational institutions and authorities use these fragments to examine speech comprehension of foreigners learning Dutch. The tests consist of 112 speech fragments, which are grouped into five difficulty levels (from 1 to 5). Good command of the Dutch language at Level 4 indicates that one is able to follow well-articulated discussions and radio and television broadcasts presented at normal speaking rates. Each level consists of 12 to 23 speech fragments and 25 four-alternative forced choice (4-AFC) questions. Most of the fragments contain read dialogues or interviews with different speakers, both male and

female. The length of the fragments varies considerably, from about a single sentence (6 words) to about 3.5 min (550 words), with a mean length of 70 s (i.e., about 160 words). The duration of the fragments increases with higher test levels.

Automatic recognition of the narratives. First, all 112 available speech fragments were recognized offline by the Dutch TNO ASR system (Human Factors, Soesterberg, The Netherlands) to generate the captions as used in this study. The TNO ASR system uses a statistical *N*-gram ($N = 3$) language model (LM) with the *N*-gram model using the $(N - 1)$ preceding words to predict the probability of the next word. The LM was trained on approximately 100 million words of newspaper texts. In total, 43 context-dependent acoustic models were trained on broadcast-news speech material, consisting of 11,880 utterances (238,724 words). The vocabulary of the ASR system (i.e., the list of recognizable words and corresponding pronunciations) consisted of 20,000 words. Words not present in the vocabulary cannot be recognized and will result in recognition errors. The ASR output was compared to the actual spoken words. We calculated the word accuracy per fragment (WA), which was defined as 100%, minus the percentage of recognition errors per fragment (Zekveld et al., 2008, in press). The WA ranged from 5% to 85%, with a mean WA of about 50%.

Selection of the narratives (Telephone-NC and interval-Telephone-NC tests). Based on the WA, the test difficulty level, and the length of the fragments, we selected six fragments to be used in the Telephone-NC in babble-noise tests. To minimize variability in test difficulty between the conditions, we selected fragments from one test level (Level 4). In the current study, the duration of the fragments had to be at least 40 s (about 100 words) to enable the participants to evaluate the benefit obtained from the captions (Möller, 2000). The ASR accuracy level of the automatic recognition of telephone conversations by near-future ASR systems will be around 70% (Duchateau, Van Uytsel, Van Hamme, & Wambacq, 2005; Fiscus et al., 2004; Stouten, Duchateau, Martens, & Wambacq, 2006). We aimed to present captions with realistic ASR accuracy levels, and therefore, the WA had to be at least 60%. Six fragments were selected, of which three fragments had a mean WA of 60.1% (range = 58% to 62.5%) and three fragments had a mean WA of 70.5% (range = 69.2% to 71.4%). The length of the fragments (mean = 150 words, range = 107 to 203 words) did not differ between the two ASR accuracy levels, to prevent confounding the effects of WA and test length.

We additionally presented two "interval-Telephone-Narrative Comprehension in Babble Noise" (Interval- Telephone-NC in babble noise) tests in which the presentation of the stimulus was interrupted after each *interval* of two sentences to ask the participants to rate the speech comprehension benefit obtained from the captions and to estimate their performance level. For these two tests, we selected two relatively long narratives from Level 4. The test consisted of 26 intervals, the WA per interval ranged from 43.5% to 87.5%, with a mean WA of 63.2%.

The long-term frequency spectrum and average root mean square power of all fragments were adapted to the sentences used in the Telephone-SRT in babble-noise tests (Plomp & Mimpen, 1979). This made the intelligibility of the narratives and SRT sentences comparable at similar SNRs. For each participant, two babble-noise levels were used in the Telephone-NC in babble-noise tests, either 54 dB SPL (*low*) and 60 dB SPL (*high*) or 64 dB SPL (*low*) and 70 dB SPL (*high*); see Procedure section.

Visual presentation of the captions (ASR output). Most ASR systems use silence detection to identify word boundaries in the incoming speech stream. As soon as a silent interval is detected, the preceding speech is processed. The length of the speech sample that is concurrently processed thus depends on the number and length of the pauses in the speech. In the current study, we aimed to simulate a realistic, though optimal, ASR system in terms of the text delay. First, we segmented the speech files using timing information of the speech pauses obtained from the TNO ASR system. Words uttered in between pauses (minimum duration of 100 ms) were grouped. The textual ASR output corresponding to those words was presented together, 100 ms after the start of the silent interval following that *speech segment*. The mean delay of displaying each word relative to the end of the word utterance was 1.1 s (range = 0.1 to 8.9 s). In online speech recognition, most ASR systems require some extra processing time to generate the text. By not including this additional delay, we simulated optimal online ASR in terms of recognition delay.

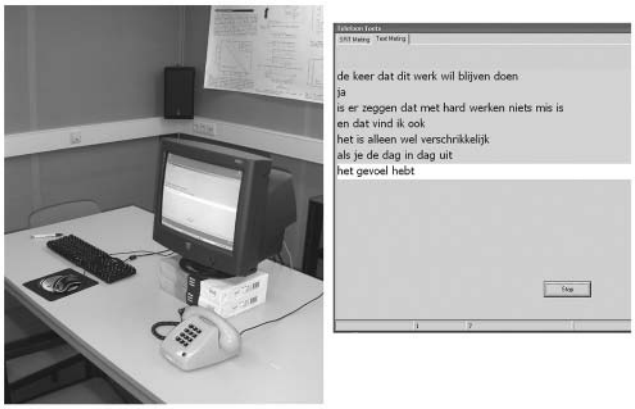

Figure 2. The left-hand panel shows the test configuration. The participants were seated in front of the PC screen and the experimenter sat at their left, behind the keyboard. Target speech signals were presented by telephone, babble noise was presented by two loudspeakers located in the corners of the test room, of which one is visible in this picture. The right-hand section shows a screen shot illustrating the presentation of the captions. The lowest text line displays the latest text output.

The text corresponding to each speech segment was presented at once, at a fixed text line. The right panel of Figure 2 illustrates the text presentation. The bottom line of the field showed the latest text output, and each new line was presented by scrolling up the preceding text with one line. In total, seven lines were presented on the screen, the first one disappearing when the eighth one was displayed. The typeface of the text was Tiresias PCfont with a font size of 16, the field background color of the eighth line was white, and the field background color of Lines 1 to 7 was green with the text being black.

Prior to the Telephone-NC in babble-noise tests, participants were instructed that they would hear a relatively long speech fragment and that partly incorrect captions would be displayed slightly delayed relative to the speech. Participants were instructed to try to comprehend the narratives with the possibility to read the captions to supplement the incomprehensible parts of the speech.

Outcome measures of the Telephone-NC tests. The outcome measures of the Telephone-NC in babble-noise test consisted of the NASA-TLX, additional rating subscales that evaluated the captions, and the performance on three 4-AFC questions on the content of the narratives.

NASA-Task Load Index. Immediately after completing each audio only or audio + text Telephone-NC in babble-noise test, participants completed the multidimensional NASA-TLX (Hart & Staveland, 1988). This paper-and-pencil scale measures the effort required to perform a task. It contains six bipolar subscales ranging from 0 to 10, which are divided into 20 increments (see Appendix A). The physical activity subscale was omitted (Zhang & Luximon, 2005). The NASA-TLX subscales that were used measure (1) mental demand, (2) time demand, (3) effort, (4) performance, and (5) frustration level. The standard NASA-TLX contains an evaluation phase (pairwise comparison procedure) in which the participants indicate the relative importance of each NASA-TLX dimension for the given task. Based on this procedure in the original version of the NASA-TLX, the subscale ratings are individually weighted to obtain the overall score. However, results of studies performed by Hill et al. (1992) and Nygren (1991) demonstrate that omitting the weighting procedure does not invalidate the workload measure (cf., Graham & Carter, 2000; Mayes, Sims, & Koonce, 2001; Moroney, Biers, & Eggemeier, 1995; Moroney, Biers, Eggemeier, & Mitchell, 1992; Murata, 1999). We decided not to administer this evaluation phase to reduce the duration of the test session. In the current study, the mean NASA-TLX (with each subscale receiving equal weight) provides the overall workload. The performance rating is inverted in the analyses; lower ratings indicate more positive (better) evaluations.

Automatic Speech Recognition-output evaluation subscales. After each Telephone-NC in babble-noise test, participants also evaluated several text features using three specific subscales similar to the NASA-TLX subscales (see Appendix A). The first subscale asked them to indicate whether the captions supported speech comprehension. The subscale ranged from "the text made speech comprehension harder" to "the text facilitated speech comprehension." The scale will be referred to as the *Text-support rating*. The second subscale asked participants to what extent the ASR errors were problematic for text comprehension (*Problems errors*), and the third subscale asked them to what extent the text delay was problematic (*Problems text delay* [not problematic . . . severely problematic]). A final 2-AFC (yes/no) question asked participants whether they would like to use the system (a device displaying textual ASR

output) in their daily lives when the amount of ASR errors and the text delay would be similar to the captions presented in the current test.

After each interval of two sentences in the interval-Telephone-NC in babble-noise test, the presentation of the speech and text was interrupted and participants were asked to complete the NASA-TLX performance subscale and the Text-support subscale.

4-Alternative Forced Choice questions. Finally, participants answered three 4-AFC questions on the content of the speech fragments. For each segment, one or two questions were available; the additional questions were constructed similar to the available ICE questions. The questions covered both the main topic and specific, detailed information provided in the narrative. The 4-AFC questions encouraged participants to really try to comprehend the spoken information. The data from the questions was not used in the analyses, as the questions were not validated for the current application.

Text Reception Threshold Test

The TRT is defined as the percentage of unmasked text required by the participant to read 50% of masked sentences without error (Zekveld et al., 2007). The sentences (Plomp & Mimpen, 1979) were displayed one by one and the text was adaptively masked with vertical bars to estimate the TRT. In each trial, the bar pattern consisted of bars of equal width. Between trials, the percentage of unmasked text was varied by changing the bar width. At the start of each trial, the mask became visible and the text appeared "behind" it in a word-by-word fashion. The timing of the appearance of each word was equal to the timing of each word in the original audio file. The preceding words remained on the screen until the sentence was completed. After the last word of the sentence was displayed, the sentence remained visible for 3.5 s. The adaptive procedure applied in the TRT test was similar to the procedure applied in the SRT test (Zekveld et al., 2007). The first sentence started with a percentage of unmasked text below threshold and this sentence was repeatedly presented with an increased percentage of unmasked text until participants were able to read the sentence correctly. The TRT is the average percentage of unmasked text for Sentences 5 to 14. Lower thresholds indicate better performance.

Spatial Span Test

The Spatial Span (SSP) test is a subtest of the CANTAB (Owen et al., 1990). It is a computerized version of the Corsi blocks task that examines visual working memory capacity (Vandierendonck, Kemps, Fastame, & Szmalec, 2004). A visual measure of spatial working memory capacity was used to prevent confounding working memory performance by hearing loss (cf., Van Boxtel et al., 2000). Another reason for applying the SSP test was to have a working memory test that does not rely on the language processes relevant for performing the TRT test (i.e., the ability to complete partly masked, written sentences).

Nine white squares were displayed at fixed pseudo-random positions on the touch-sensitive screen. Several squares changed color sequentially in each trial. The number of squares that changed color per trial increased from two to a maximum of nine. After a tone prompt, participants had to touch the squares that had changed color, in the correct order. If they repeated the sequence incorrectly, an alternative sequence of the same length was presented. The test ended automatically if participants failed at three consecutive trials of one level. The SSP was the highest sequence of squares recalled successfully.

Questionnaire

At the end of the test session, the participants completed a questionnaire. Part 1 focused on speech comprehension problems during daily telephone communication, and Part 2 evaluated the captions and the benefit obtained from the text.

Pilot Study

Because the babble noise and the experimental set-up were specifically prepared for the current study, a pilot study was performed to estimate the Telephone-SRT in babble noise and subjective intelligibility of the narratives in the Telephone-NC in babble-noise tests for normal hearing listeners. Seven normal hearing participants (mean age = 27 years; two males) performed four Telephone-SRT tests in babble-noise with a level of 60 dB SPL. The data of the first (practice) Telephone-SRT in babble-noise test were not used in the analysis. The mean Telephone-SRT in babble noise was −10.9 dB SNR and ranged from −7.5 to −14.3 dB SNR. The participants also performed Telephone-NC tests in babble-noise at several SNRs and were asked to indicate the

listening effort and to estimate their speech comprehension performance. Based on the results, we decided to set the default SNR in the Telephone-NC tests in babble noise at 0 dB for the hearing-impaired participants. Note that for most of them, this default level was adapted, depending on the individual Telephone-SRT in babble noise (see Procedure section).

Procedure

The test procedure aimed to approximate daily telephone conversations in the presence of babble noise. At the start of the test session, participants were asked which ear they normally used during telephone conversations and, if appropriate, whether they used a hearing aid, a special hearing aid program, or the telecoil setting of their hearing aid. They were instructed to perform the tests with the same ear, hearing aid settings, and position of the telephone receiver normally used during telephone conversations. They were not allowed to change these settings during the test session.

First, the participants performed a practice Telephone-SRT test in quiet, followed by another Telephone-SRT test in quiet used to determine the level of the babble noise in the Telephone-SRT and Telephone-NC tests. If the Telephone-SRT in quiet was below 45 dB, the *high* babble-noise level in the Telephone-SRT and Telephone-NC tests was 60 dB SPL; otherwise, the *high* babble-noise level was 70 dB SPL. Higher babble-noise levels were not presented because this would result in unrealistically high babble-noise and speech levels in the Telephone-SRT and Telephone-NC tests. Then, a practice Telephone-SRT in babble-noise test was presented to make the participant familiar with the test, followed by two other Telephone-SRT in babble-noise tests, the results of which were used in the analysis. Then they performed a practice Telephone-NC in babble-noise test and completed the NASA-TLX, the ASR-output evaluation subscales, and practice 4-AFC questions. After that, they performed three actual Telephone-NC in babble-noise tests, followed by three TRT tests and the interval-Telephone-NC in babble-noise test. Then they performed the remaining three Telephone-NC in babble-noise tests, two Telephone-SRT in babble-noise tests, three TRT tests, and the SSP test. The 1.5-hr test session was finished after completing the questionnaire. Sentences presented in the Telephone-SRT in babble-noise and TRT tests were only presented once to each participant.

In total, six Telephone-NC in babble-noise tests were performed. In half of the tests, the high babble-noise level was used (60 or 70 dB SPL), and in the other tests, the low babble-noise level (54 or 64 dB SPL) was used. This enabled us to examine interaction effects between the noise level and the subjective benefit obtained from the captions. The difference between the Telephone-SRT in babble-noise and the SNR applied in the Telephone-NC in babble-noise tests will be referred to as the *relative speech intelligibility level*, with more positive values reflecting better speech intelligibility. Depending on the individual Telephone-SRT in babble noise, the speech level in the Telephone-NC in babble-noise tests was adapted; if the Telephone-SRT in babble noise (mean of the first two tests) exceeded ±6 dB SNR, the speech level was increased or decreased with 6 dB steps until the relative speech intelligibility level was between −6 and +6 dB SNR. This ensured that all participants were able to comprehend part of the speech. We assumed that presenting the speech in the tests at a SNR far below the individual Telephone-SRT in babble noise would not reflect daily life telephone conversations. In reality, listeners would probably improve the listening conditions in such difficult situations by, for example, trying to reduce the background noise. Furthermore, for some participants with a relatively good Telephone-SRT in babble noise, the SNR was adapted to prevent ceiling effects (i.e., easy speech comprehension, which could reduce the subjective benefit obtained from the captions). In two Telephone-NC in babble-noise tests, no ASR output was presented. This resulted in a 2×3 design: two babble-noise levels (high or low) by three ASR conditions (no captions, captions with mean WA of 60%, captions with a mean WA of 70%). Across participants, each Telephone-NC in babble-noise test was presented equally often in each order position and each fragment was presented about equally often in each condition. This prevented the test condition from becoming confounded with the test order and speech fragment.

Results Experiment 1

Results of the Speech Reception Threshold, Text Reception Threshold, and Spatial Span Tests

The mean Telephone-SRT in quiet was 50.8 dB SPL, (*SD* = 12.2 dB) and the mean Telephone-SRT in babble noise was 2.9 dB SNR (*SD* = 10.0 dB).

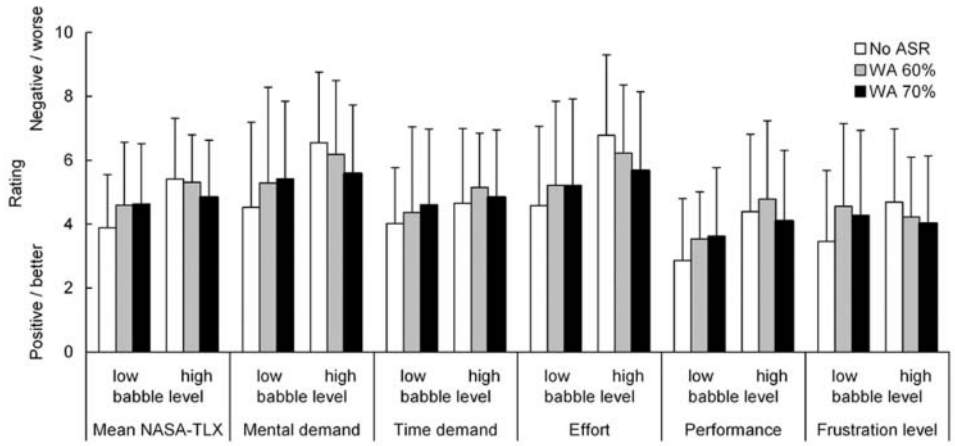

Figure 3. Results on the NASA-Task Load Index (NASA-TLX). The mean task load is shown for each of the six conditions at the left-hand side of the figure. The other bars show the mean subscale ratings. Lower values indicate better, or more positive, ratings. ASR = automatic speech recognition; WA = word accuracy.

The mean TRT was 51.7% (*SD* = 5.6%) and the mean SSP (visual working memory capacity) was 5.3 (*SD* = 1.3).

Results of the Telephone-Narrative Comprehension in Babble-Noise Tests

For 17 participants, the speech level in the Telephone-NC in babble-noise tests was adapted based on the individual Telephone-SRT in babble noise. The mean relative intelligibility level was 2.24 dB SNR (*SD* = 3.7 dB). Figure 3 shows the results of the NASA-TLX subscales. From left to right, the first six bars show the mean NASA-TLX rating for each of the six conditions. Lower values indicate better, or more positive, ratings.

We aimed to test the main and interaction effects of ASR condition (no ASR, WA 60%, and WA 70% WA) and babble-noise level (high vs. low) on the ratings on the NASA-TLX and the ASR-output evaluation scales. Applying nonparametric statistical tests is recommended for the analysis of nonnormal and ordinal-scales data (Altman, 1991; Svensson, 2001). However, inspection of the rating scale data revealed that for each variable, at least 70 different data values occurred in the data set, and not many identical data values were observed. Tests for skewness and kurtosis did not indicate a nonnormal distribution of the data. Given these results and the absence of a powerful nonparametric test for a two-within factors repeated-measures analysis, we decided to use (parametric) univariate repeated-measures analyses.

The subjective NASA-TLX ratings were analyzed with univariate repeated-measures analyses with two within-subjects variables: ASR condition (no ASR, WA 60%, and WA 70%) and babble-noise level (high or low). Simple contrasts (with Bonferroni adjustments) were used to test for differences between the no-ASR condition and either the 60% or the 70% WA condition. Degrees of freedom were adjusted (Greenhouse–Geisser correction) if Mauchly's test of sphericity indicated violation of the sphericity assumption.

The repeated-measures analyses indicated that the main effect of ASR condition was not statistically significant for the mean NASA-TLX rating and the ratings on the individual subscales. This indicates that the task load did not differ between the no-ASR condition, the 60% WA condition, and the 70% WA condition. The interaction effect between babble-noise level and ASR condition was statistically significant for the mean NASA-TLX rating ($F(2, 38) = 4.72$; $p = .015$), the effort subscale ($F(2, 38) = 5.51$; $p = .008$), and the frustration level subscale ($F(2, 38) = 5.81$; $p = .006$). The main effect of babble-noise level was statistically significant for the mean NASA-TLX rating ($F(1, 19) = 8.15$; $p = .010$), the mental demand subscale ($F(1, 19) = 8.42$; $p = .009$), the effort subscale ($F(1, 19) = 11.5$; $p = .003$), and the performance subscale ($F(1, 19) = 14.2$; $p = .001$). None of the contrasts between the no-ASR condition and the 60% and 70% WA levels was statistically significant ($p > .05$). These results indicate that the NASA-TLX ratings did not differ between

Table 2. Means and Standard Deviations (Between Parentheses) of the ASR-Output Evaluation Subscales,[a] and the Percentage of Participants Who Indicated That They Would Like to Use the System in Their Daily Lives

	Low Babble-Noise Level		High Babble-Noise Level	
	WA 60%	WA 70%	WA 60%	WA 70%
Text support (support . . . hindrance)	5.6 (2.2)	4.5 (1.8)	6.2 (1.5)	5.6 (2.4)
Problems with ASR errors (not problematic . . . severely problematic)	5.4 (2.8)	4.8 (2.5)	5.6 (2.4)	4.9 (2.2)
Problems with text delay (not problematic . . . severely problematic)	4.8 (2.8)	4.2 (2.7)	5.1 (2.5)	4.4 (2.5)
Willingness to use the system (% of participants)	25	15	10	15

NOTES: ASR = automatic speech recognition; WA = word accuracy.
a. The subscales ranged from 0 to 10; lower ratings reflect better evaluations of the captions.

the no-ASR conditions and the conditions in which captions were presented. The interaction effect between babble-noise level and ASR condition indicates that for the low babble-noise levels, presenting the captions increased the mean task load, the effort, and the frustration level, whereas for the high babble-noise levels, presenting the captions decreased the mean task load, effort, and frustration level (Figure 3). The main effect of babble-noise level indicates that for the high babble-noise levels, the mean task load, mental demand, effort, and performance ratings were higher (more negative) than for the low babble-noise levels.

The means and standard deviations (between parentheses) of the results of the ASR-output evaluation subscales are shown in Table 2. Lower values reflect better evaluations of the captions. The results on the ASR-output evaluation scales were analyzed with a univariate repeated-measures analyses with two within-subjects variables WA level (WA 60% or WA 70%) and babble noise-level (high or low). Note that in the no-ASR conditions, the participants did not complete the ASR-output evaluation scales.

The repeated-measures analyses showed that for the ASR-output evaluation scales, the interaction effect between babble-noise level and WA level, and the main effect of WA were not statistically significant. The main effect of babble-noise level reached statistical significance for the "problems with text delay" rating ($F(1, 19) = 4.65$; $p = .044$). This effect indicates that the delay of the text is more problematic when the babble-noise level is high.

The mean percentage of participants who indicated that they would like to use the system is shown in Table 2. We used nonparametric sign tests to test for the effect of WA on the mean percentage of participants indicating that they would like to use the system (tested separately for each babble-noise level). The sign tests (with Bonferroni adjustments) indicated no statistically significant effect of WA on the willingness to use the system ($p > .10$).

Interval-Telephone-Narrative Comprehension in Babble-Noise Tests

Table 3 shows the results of the interval-Telephone-NC in babble-noise tests. We calculated the Spearman correlation coefficients between the mean WA per subfragment and the two subscale ratings. The Spearman correlation coefficient between the mean text-support rating and WA was statistically significant ($r = -.52$, $p < .01$). As can be seen in Table 3, the mean text-support rating was 6.3, implying that for most participants, the text made speech comprehension more difficult. The correlation between WA and the text-support rating indicates that this "hindrance" from the text was lower for higher ASR accuracies.

Practice Effects

Participants performed six Telephone-NC in babble-noise tests in the test session. Participants may become more experienced with using the captions during the course of the test session, which could be reflected in the outcome measures. In order to test for practice effects on the results of the Telephone-NC in babble-noise test, we performed repeated-measures analyses with order position as independent variable, and either the mean NASA-TLX rating or one of the three ASR-output evaluation ratings as dependent variable. The analyses indicated no order effects on the dependent variables ($p > .10$).

Table 3. Means, Standard Deviations (Between Parentheses), and Range of the Ratings Obtained in the Interval-Telephone-Narrative Comprehension in Babble-Noise Tests[a]

		Range	
	Mean (*SD*)	Minimum	Maximum
Performance (high . . . low)	4.9 (0.6)	4.0	6.0
Text support (support . . . hindrance)	6.3 (0.7)	4.9	7.4

a. The subscales ranged from 0 to 10; lower values reflect better evaluations.

Table 4. Spearman Correlation Coefficients Between Several Individual Variables and the Outcome Measures of the Telephone-Narrative Comprehension in Babble-Noise Tests[a]

	Mean NASA-TLX		Text Support (Support . . . Hindrance)	Problems With ASR Errors (Rating)	Problems With Text Delay (Rating)	Willingness to Use System (% of Participants)
	No ASR	ASR				
Age	−.25	−.53*	−.16	−.45*	−.31	.23
RSIL	−.27	−.07	.29	−.23	.13	−.24
TRT	−.12	−.15	.05	.10	−.04	.37
SSP	.28	.20	.22	−.15	−.03	−.48*

NOTES: NASA-TLX = NASA Task Load Index; ASR = automatic speech recognition; RSIL = relative speech intelligibility level; TRT = text reception threshold; SSP = spatial span.

a. The rating subscales ranged from 0 to 10. Higher SSP values reflect higher working memory capacities. Higher values on the subscales reflect higher task load or worse evaluations, higher TRTs reflect worse performances.

*$p < .05$.

Correlation Analysis

In a correlation analysis, we examined whether age, the relative speech intelligibility level, the TRT, and the SSP were associated with the reported task load and the evaluation of the captions. The strength of those relationships may depend on whether text was presented or not. We therefore separately calculated the Spearman correlation coefficients for the ASR and no-ASR conditions (Table 4).

The Spearman correlation coefficient between age and the mean NASA-TLX in the conditions in which the captions were presented was statistically significant ($r = -.53$; $p < .05$), indicating an age-related decrease in the reported task load. A significant correlation between age and the "problems with the ASR errors" rating consistently indicated that the older participants evaluated the ASR errors as less problematic than the younger participants did ($r = -.45$; $p < .05$). Finally, the participants with larger working memory capacities (SSPs) indicated less willingness to use the system ($r = -.48$, $p < .05$). Inspection of the raw data, however, revealed that this relationship was due to one participant having a relatively small working memory capacity.

The results on the questionnaire are presented together with the results of the participants of Experiment 2 in Appendix B.

Discussion Experiment 1

The main result of Experiment 1 was that hearing impaired participants indicated that it is difficult to obtain speech comprehension benefit from ASR output with mean ASR WA levels of 60% or 70%. The task load did not differ between the conditions in which no ASR output was presented and the conditions in which ASR output with WA levels of 60% or 70% were presented. However, the interaction between ASR condition (captions absent, WA 60%, or WA 70%) and babble-noise level (low or high) indicated that the overall task load decreased when captions were presented, but only for the high babble-noise level. For the low babble-noise level, presenting the text seemed to increase the task load (Figure 3). The analyses of the subscale ratings indicated that this interaction effect was mainly based on the effort and frustration level subscales. Not surprisingly, the task load was generally higher for the

high babble-noise level than for the low babble-noise level.

In our earlier study (Zekveld et al., in press) we concluded that hearing impaired participants obtained benefit from partly incorrect ASR output in the comprehension of speech in noise. The ASR accuracy level of the text presented in that study ranged from 37% to 74%; the SRT improved by about 1.5 to 2 dB SNR. The participants of the current study, however, reported that captions with WAs of 60% and 70% do not decrease the task load during speech comprehension in difficult listening situations. Several factors could explain the apparently inconsistent results. First, the SRT tests applied in the Zekveld et al. (2008, in press) studies substantially differed from the Telephone-NC_{BB} tests used in the current study. In our previous studies, participants had time to listen to the sentence and subsequently read the corresponding text before the next sentence was auditorily presented. In the current study, participants had to listen to ongoing speech while reading captions corresponding to previously uttered speech, which likely made it more difficult to combine the speech and the captions. It is widely accepted that it is difficult to perform two verbal tasks simultaneously, even when the verbal stimuli are audiovisually presented (cf., Bourke, Duncan, & Nimmo-Smith, 1996; Jobard, Vigneau, Mazoyer, & Tzourio-Mazoyer, 2007). The "perceptual overload" in the Telephone-NC in babble-noise tests was furthermore illustrated by the remarkably similar comments of the participants: most of them spontaneously and carefully explained that they found it difficult to read the captions while listening to the ongoing speech. The ASR errors distracted them from listening and made text comprehension problematical. As described by Pichora-Fuller, Schneider, and Daneman (1995), more working memory capacity is required when speech or text comprehension is difficult, for example, when background noise masks the speech. The current results could suggest that for the low babble-noise levels, the potential subjective benefit obtained from the captions does not compensate for the increment in working memory capacity required for processing the partly erroneous text (Yeh & Wickens, 1988). In other words, the "costs" (increments in the subjective task load) associated with processing the captions may not be compensated for by better speech comprehension. The interaction effect between ASR condition and babble-noise level may reflect a shift in this balance when speech comprehension itself is more difficult and effortful, thereby increasing the potential support obtained from the text.

Second, objectively measured benefit does not have to be consistent with subjectively experienced benefit (e.g., Möller, 2000; Saunders, Forsline, & Fausti, 2004; Wickens, 1992). Consistently with the task load increment when captions were presented when the level of the babble-noise was low, in our previous study (Zekveld et al., in press), presenting the text increased the effort when the text was delayed relative to the speech, despite objective speech comprehension benefit in these conditions.

The current data do not allow a conclusion regarding the cause of the apparent difficulties of hearing impaired participants to use ASR output to improve speech comprehension. Captions with accuracy levels exceeding 70% may reduce the task load and additionally, presenting the text prior to the corresponding speech could make it easier to use the captions during speech comprehension. To examine whether the limited ability of the participants to use the text was caused by the number of ASR errors, the delay of the text relative to the speech, or both, a follow-up experiment was performed.

Experiment 2

In Experiment 2, we added two conditions with high ASR accuracy levels (WAs of 80% and 90%) and in half of the conditions, the text was presented prior to the utterance of the corresponding speech, which allowed participants to read along with the speech. Note that in actual ASR applications, the text will not precede the speech, unless the speech is artificially delayed.

Methods Experiment 2

Unless stated otherwise, the methods and procedure applied in Experiment 2 were equal to those in Experiment 1.

Participants

A total of 10 hearing-impaired patients (5 female, 5 male) were tested in Experiment 2; none of them had participated in Experiment 1. Their ages ranged from 45 to 74 years, with a mean age of 62.7 years

(*SD* = 9.9 years). Most participants had sensorineural hearing loss, but 3 participants had mixed sensorineural/conductive hearing loss. Means and standard deviations of the pure-tone hearing thresholds (average of both ears) are shown in Figure 1. In all, 5 participants used two hearing aids, and 1 participant used one hearing aid in daily face-to-face conversations. Only one of them used a hearing aid in daily telephone communication (without telecoil setting or special hearing aid program), and 2 participants used telephone amplifiers.

Stimuli and Tests

Telephone-Narrative Comprehension in Babble-Noise Tests

Instead of the manipulation of the babble-noise level, in Experiment 2, we varied the text delay: The text either preceded or followed the corresponding speech. Four additional narratives were required for two extra WA levels (80% and 90%) by two delay conditions (lead vs. lag). Because no narratives were available with WAs exceeding 70% that were comparable with the narratives presented in Experiment 1, we first selected four narratives with similar test difficulty and length as the narratives of Experiment 1, and then randomly corrected the ASR errors to obtain two fragments with WA levels of 80% and two fragments with WA levels of 90%. Each participant performed 10 Telephone-NC in babble-noise tests: in two tests, no captions were presented, and the WA of the captions presented in the remaining eight conditions was 60%, 70%, 80%, or 90%. For each WA level, the captions either preceded or followed the speech. In contrast to Experiment 1, the interval-Telephone-NC in babble-noise test was not presented, to reduce the length of the test session. Participants of Experiment 2 were not asked to complete the "problems with the text delay" rating scale, as in half of the tests, the text preceded the speech.

Visual Presentation of ASR Output

In the *lag* conditions, the timing of the presentation of the text relative to the speech was equal to Experiment 1. In the *lead* conditions, the onset of the visual presentation of the ASR output for each segment was presented at the *start* of the utterance of the first word of that segment. This means that in the lead conditions, the text was available prior to the pronunciation of the corresponding words and participants were able to read the text along with listening to the corresponding speech.

Procedure

The order of the tests was the same as in Experiment 1, except that 10 Telephone-NC in babble-noise tests were presented in two blocks of five tests each. Similar to Experiment 1, for some participants, the levels of the speech and babble-noise were adapted depending on the individual Telephone-SRT_{BB}.

Results Experiment 2

Results of the Speech Reception Threshold, Text Reception Threshold, and Spatial Span tests

The mean Telephone-SRT in quiet was 55.2 dB SPL (*SD* = 13.8) and the mean Telephone-SRT in babble noise was −1.1 dB SNR_{BB} (*SD* = 7.1 dB). The mean TRT was 53.1% (*SD* = 5.55 %) and the mean SSP (visual working memory capacity) was 4.7 (*SD* = 1.0). These results were similar to the results of the participants of Experiment 1.

Results of the Telephone-Narrative Comprehension in Babble-Noise Tests

For 6 participants, the speech level in the Telephone-NC in babble-noise tests was adapted based on the individual Telephone-SRT in babble noise. The mean relative intelligibility level was 2.29 dB SNR (*SD* = 1.2 dB). Figure 4 shows the NASA-TLX results.

First, we used a repeated-measures analysis to test the main and interaction effects of WA condition (60%, 70%, 80%, or 90% WA) and delay (lead vs. lag) on the ratings on the NASA-TLX and the ASR-output evaluation scales. Simple contrasts (with Bonferroni adjustments) were used to test for differences between the 70%, 80%, and 90% WA conditions relative to the 60% WA condition. Similar to Experiment 1, based on the high number of different data values for each variable and the results of tests for skewness and kurtosis, we decided to use parametric repeated measures analyses. Note that the results of the analyses should be interpreted with caution, as the sample-size of the current experiment was small (i.e., *N* = 10). Degrees of freedom were adjusted (Greenhouse–Geisser correction) if

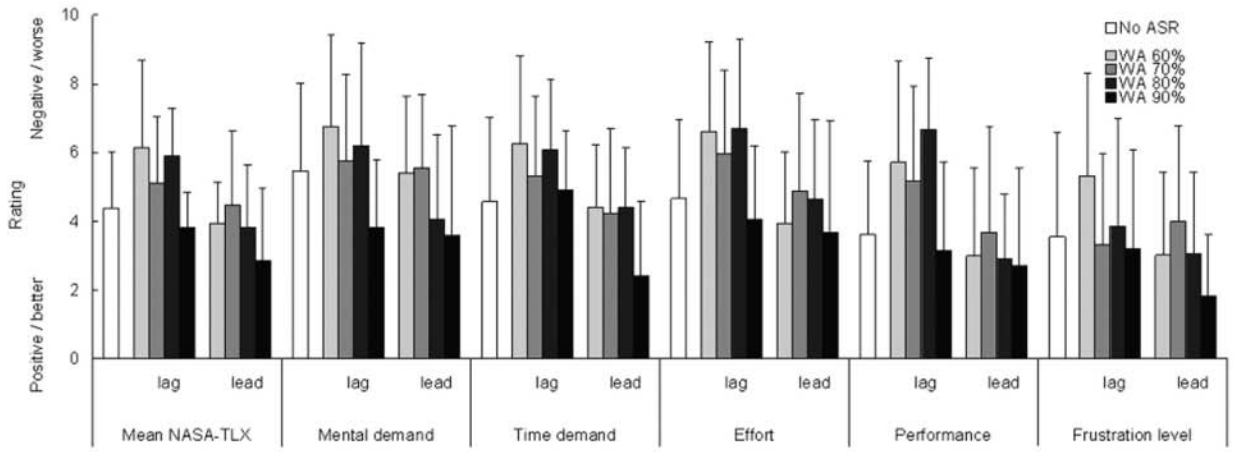

Figure 4. Mean results of the NASA-Task Load Index (NASA-TLX). The mean task load is shown for each of the six conditions in the left-hand side of the figure. The other bars show the subscale ratings. Lower values indicate better, or more positive, ratings. ASR = automatic speech recognition; WA = word accuracy.

Mauchly's test of sphericity indicated violation of the sphericity assumption.

The repeated measures analyses indicated main effects of WA on the mean NASA-TLX rating ($F(2, 27) = 6.237$; Greenhouse–Geisser corrected p value = .009; $\varepsilon = .65$), the time demand subscale ($F(3, 27) = 9.15$; $p < .001$), the effort subscale ($F(3, 27) = 3.04$; $p = .046$), and the performance subscale ($F(3, 27) = 3.94$; $p = .019$). The main effect of delay was statistically significant for the mean NASA-TLX rating ($F(1, 9) = 10.30$; $p = .011$), the effort subscale ($F(1, 9) = 13.98$; $p = .005$), and the performance subscale ($F(1, 9) = 11.19$; $p = .009$). For none of the subscales, the interaction effect between WA and delay was statistically significant. The results indicate that for higher WA levels, the mean task load and the time demand, effort, and performance ratings are lower (more positive). Additionally, the mean task load and the effort and performance ratings are also lower (better) for the conditions in which the text is presented prior to the corresponding speech. Thus, both increasing the WA level of the ASR output, and omitting the delay of the text relative to the speech, reduces the task load experienced by the listeners. The contrast between the 60% WA and the 90% WA conditions was statistically significant for the mean NASA-TLX rating ($F(1, 9) = 23.1$; Bonferroni corrected $p = .003$), the mental demand subscale ($F(1, 9) = 40.0$; Bonferroni corrected $p < .001$), and the time demand subscale ($F(1, 9) = 15.05$; Bonferroni corrected $p = .012$). Consistent with the main effect of WA level, these results indicate that increasing the accuracy of the text from 60% to 90%, reduces the subjective task load, mental demand, and time demand, as averaged over the lead and lag conditions.

Note that the no-ASR condition was not included in the repeated-measures analysis, as this condition was not crossed with the delay factor (the lead/lag distinction does not apply for this condition as no text was presented). Therefore, the data analysis contained a second step in which we performed two repeated measures analyses, each with one within-subject variable "ASR condition." In one analysis, the three ASR-condition levels were no-ASR, WA 60% *lag*, and WA 60% *lead*, and in the second analysis, the levels were no-ASR, WA 90% *lag*, and WA 90% *lead*. The rationale for only analyzing these "extreme" WA conditions was that the 60% WA conditions simulated *realistic* ASR performance, whereas the 90% WA conditions simulated *optimal* (future) ASR performance. The 70% and 80% WA conditions were not included in this analysis to reduce the number of statistical tests conducted on the small data set.

For the repeated-measures analyses with the within-subject variable ASR-condition with levels no-ASR, 60% WA lag, and 60% WA lead, the effect of ASR-condition was statistically significant for the mean NASA-TLX rating ($F(2, 18) = 3.68$; $p = .046$) and the effort rating ($F(2, 18) = 3.95$; $p = .038$). As can be seen in Figure 4, these results are based on the relative high mean NASA-TLX and effort ratings in the 60% WA conditions in which the text was delayed relative to the speech. The repeated-measures analyses on the results of the no-ASR, 90% WA lag, and 90% WA lead conditions indicated a main effect of

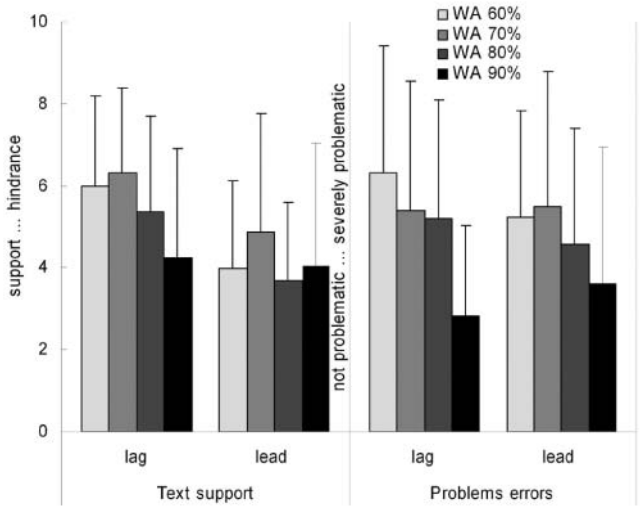

Figure 5. Means and standard deviations (error bars) of the ratings on the automatic speech recognition-output evaluation scales. The subscales ranged from 0 to 10; lower values reflect better evaluations. WA = word accuracy.

ASR condition for the time demand subscale ($F(2, 18) = 4.00$; $p = .037$). Figure 4 shows that presenting ASR output with WA of 90% slightly increased the time demand when the text was delayed relative to the speech, but presenting the text prior to the speech reduced the time demand compared with the conditions in which no ASR output was presented. These results indicate that for WA levels around 60%, the task load is relatively high when the text is delayed relative to the speech. Simulating optimal ASR accuracies (WA around 90%) and presenting the text prior to the speech reduces the subjective time demand ratings as compared with the conditions in which no ASR output was presented. Unfortunately, in real online ASR applications, the text will not be available prior to the speech.

Means and standard deviations of the ASR-output evaluation ratings are shown in Figure 5. Lower ratings reflect better evaluations.

We performed repeated-measures analyses to test the effects of the within-subject variables WA condition (60%, 70%, 80%, or 90% WA) and delay (lead vs. lag) on the text-support ratings and the "problem with errors" rating. Simple contrasts (with Bonferroni adjustments) were used to test for differences between the 70%, 80%, and 90% WA conditions relative to the 60% WA condition. The results of the analysis indicated no statistically significant effects of ASR accuracy and delay on the text-support ratings. The main effect of WA was statistically significant ($F(3, 27) = 5.10$; $p = .006$) for the "problems with errors" rating. For this ASR-output evaluation scale, the contrast between the 60% and 90% WA condition was also statistically significant ($F(1, 9) = 20.4$; Bonferroni corrected $p = .003$); see the right-hand panel of Figure 5.

The mean percentage of participants willing to use the system in daily life is shown in Table 5. We used nonparametric sign-tests to test for the effect of WA on the mean percentage of participants indicating that they would like to use the system (tested separately for each babble-noise level). The sign tests (with Bonferroni adjustments) indicated no statistically significant effect of WA on the willingness to use the system ($p > .10$).

Correlation Analysis

Spearman correlation coefficients were calculated between age, the relative speech intelligibility level, the TRT, the SSP, the reported task load, and the evaluation of the captions, separately for the ASR, no-ASR, lead, and lag conditions (Table 6).

Higher ages were significantly associated with lower reported task load (mean NASA-TLX) in the conditions in which no captions were presented ($r = -.64$, $p < .05$) and in the lead conditions ($r = -.82$, $p < .01$). Elderly participants reported fewer problems with the ASR errors in the lead conditions ($r = -.65$, $p < .05$) and a higher age was related to more willingness to use the system in daily situations ($r = .78$, $p < .01$). In the lag conditions, participants with better (lower) TRTs reported lower task load than participants with worse TRTs ($r = .79$, $p < .05$). Better TRTs were also related to more support obtained from the text in the lag conditions ($r = .68$, $p < .05$). Thus, better TRTs were associated with lower task load and better evaluations of the ASR output in the lag conditions. The performance on the TRT test was not significantly associated with age.

Discussion Experiment 2

The aim of Experiment 2 was to examine whether increasing the WA level of the captions and/or presenting the captions preceding the utterance of the speech reduced the subjective task load. The reported task load indeed decreased for the higher WA levels; participants indicated that the time demand and effort were lower for the higher WA levels,

Table 5. Mean Percentage of Participants Who Reported That They Would Like to Use the System in Their Daily Lives

	Lag Conditions				Lead Conditions			
	WA 60%	WA 70%	WA 80%	WA 90%	WA 60%	WA 70%	WA 80%	WA 90%
Willingness to use the system (% of participants)	10	10	20	40	10	20	40	40

NOTE: WA = word accuracy.

Table 6. Spearman Correlation Coefficients Between Several Individual Variables and the Outcome Measures of the Telephone-Narrative Comprehension in Babble-Noise Tests[a]

	Mean NASA-TLX			Text Support Rating (Support . . . Hindrance)		Problems With ASR Errors (Rating)		Willingness to Use System (% of Participants)	
		ASR							
	No ASR	Lag	Lead	Lag	Lead	Lag	Lead	Lag	Lead
Age	−.64*	−.20	−.82**	.19	−.55	−.38	−.65*	.13	.78**
RSIL	−.19	−.01	.13	−.04	.16	−.24	−.04	−.05	−.20
TRT	.20	.79*	.24	.68*	−.20	.43	.15	−.35	.15
SSP	−.11	.11	.29	.04	−.19	−.35	.15	.40	.46

NOTES: NASA-TLX = NASA Task Load Index; ASR = automatic speech recognition; RSIL = relative speech intelligibility level; TRT = text reception threshold; SSP = spatial span.
a. The rating subscales ranged from 0 to 10. Higher SSPs reflect larger working memory capacities. Higher values on the rating scales reflect worse evaluations, and higher TRTs reflect worse performances.
*$p < .05$. **$p < .01$.

and the subjective performance level increased for higher WA levels. The reported task load was lower in the 90% WA conditions compared with the 60% WA conditions and participants reported having fewer problems with the ASR errors in the 90% WA conditions than in the 60% WA conditions. Furthermore, presenting the text *prior to* rather than *after* the corresponding speech decreased the task load and subjective effort, and improved the performance rating. The results of the second step of the analysis indicated that in the conditions in which the text was delayed to the speech, presenting the text increased the mean task load and the effort ratings. In contrast, when the text preceded the speech, the time demand reported by the participants was lower as compared with the no-ASR conditions. Thus, increasing the accuracy of the ASR output and omitting the text delay resulted in lower task load and better evaluations of the captions in the conditions in which ASR output was presented. Only when the WA of the text was high (90%) and the text could be read simultaneously to hearing the corresponding speech, the participants indicated that one task load dimension, the time demand, was reduced compared to audio-only speech comprehension. When the ASR accuracy was around 60% and the text followed the speech, the captions did not reduce the task load, but even slightly increased the mean task load and the effort ratings. This result is inconsistent with the similar or slightly reduced overall task load and effort ratings for the same conditions as observed in Experiment 1 (see Figure 3). The different results for the two experiments could be because of the fact that the participants of Experiment 2 also performed conditions in which the WA level was high and the text was not delayed to the speech. This may have resulted in a different "frame of reference" adopted by the participants in Experiment 2, resulting in a relatively poor evaluation of the most difficult condition performed by this group (60% WA, lag). Note, however, that the current results should be interpreted cautiously as the number of participants in Experiment 2 was relatively small. The remaining findings of Experiment 2 are discussed in the General Discussion section, together with the results of Experiment 1.

General Discussion

The main result of the current study is the participants' report that automatically generated captions with word accuracies of 60% to 90% that were presented slightly delayed or preceding the corresponding speech did not reduce the task load compared to when no captions were presented. The results of Experiment 1 suggest that the processing of the visual information increased the task load in the low babble-noise conditions, whereas the effort and frustration decreased when captions were presented in the high babble-noise conditions. The participants of Experiment 2 however indicated that ASR output with 60% accuracy that was presented delayed to the speech increased the mean NASA-TLX rating and effort rating. Only the time demand ratings reduced when highly accurate captions (i.e., 90% WA) were presented prior to the speech. In general, despite reduced task load with increasing WAs, the results of both experiments suggest that the subjective task load generally did not differ between the audio-only and audio + text conditions. This indicates that the additional visual information did not substantially influence the perceived effort required to perform the listening task. Thus, despite the objective speech comprehension benefit as observed in our earlier studies (Zekveld et al., 2008, in press), the participants did not report that the text reduced the task load. Reading the partly incorrect captions is a difficult task to be performed while listening to ongoing telephone speech in the presence of babble noise in the test room. Processing the text imposes an additional task on the participants, and generally, when extra, useful information is provided to participants, the reported task load increases, even if the additional information improves the performance (Yeh & Wickens, 1988). The current study showed that the task load was not increased by the additional visual information when the accuracy of the text was higher than 60% or when the text preceded the speech. These results may suggest that in these conditions, the extra task demands by the processing of the text were partly compensated for by less effortful speech comprehension. The current results could, however, also reflect the common finding that, in contrast to objective performance measures, subjective task load measures are relatively insensitive to changes in working memory demands when the task load is high (Yeh & Wickens, 1988). Importantly, regardless of the underlying cause of the current results, the present study underlines the need for examining both subjective and objective speech comprehension benefit obtained from assistive communication systems.

Similar to the current findings, Leitch (2008) described how students who used the Liberated Learning system indicated that combining erroneous text and speech is difficult. Automatically generated captions were presented to about 44 high school students during lectures. For ASR accuracies around 65%, more than 70% of the students reported that the captions did not improve the comprehension of the lectures. More than 40% of the students indicated having great difficulties understanding the text. Much like the current results, the students reported that poor ASR accuracy levels, difficulties understanding the text, and distraction reduced the benefit obtained from the visual information and their willingness to use the system.

In Experiment 1, a higher age was associated with lower reported task load when captions were presented. This relation was also observed in Experiment 2 for the conditions in which the captions preceded the speech and for the conditions in which no captions were presented. In both experiments, elderly people reported fewer problems with the ASR errors (i.e., in Experiment 2, this relation was observed for the lead conditions). Thus, a higher age was associated with lower subjective task load and less difficulty with ASR errors. Note that in Experiment 2, a relation between age and the mean task load was also observed for the conditions in which no ASR output was presented; this lower "baseline" task load likely has contributed to the relation between age and the task load in the lead conditions.

In our previous study (Zekveld et al., in press), aging was associated with more effort required for combining the speech and the text. Specifically, a higher age was related to a larger *difference* between the effort ratings in the auditory tests and the audiovisual tests. In the current study, we analyzed the relation between age and the "absolute" NASA-TLX ratings. In both experiments of the current study, the relationship between age and the *difference* between the mean NASA-TLX ratings in the no-ASR and ASR conditions was not statistically significant. Other differences between the tasks and the subjective outcome measures applied in the current study and our previous study likely also have contributed to the different findings.

To check whether the age effects observed in the current study were caused by a mediating relationship between the other variables, we calculated the

Spearman correlation coefficients between age, the relative intelligibility level, the TRT, and the SSP. None of these correlations was statistically significant, except for a correlation between age and the relative intelligibility level in Experiment 1. After controlling for the relative intelligibility level, a higher age was significantly associated with lower reported task load, fewer problems with the delay of the text, and a higher willingness to use the system (Experiment 1). These results are quite surprising considering the relatively high complexity of the task and the fact that aging is often associated with decreased performances on auditory tests.

The results of the current study could also reflect that elderly participants tend to give more "positive" evaluations than the younger participants did. Other studies have reported that elderly persons may respond differently on questionnaires than younger persons (e.g., elderly people report greater satisfaction in health care evaluations; Fitzpatrick, 1991). However, although some studies reported age-related decreases in the NASA-TLX ratings (e.g., Tomporowski, 2003), others observed that elderly participants reported higher task load than younger participants (Deaton & Parasuraman, 1993; Graham & Carter, 2000). In Experiment 2, a higher age was additionally associated with a higher willingness of the participants to use the system in daily life. This relation could be associated with the lower task load reported by the elderly participants, but it could also reflect that younger participants associate the system with ageing, which can reduce their willingness to use the system (Southall, Gagné, & Leroux, 2006). Other variables could also influence the willingness of hearing impaired participants to use a system that automatically generates captions during telephone conversations. Evidently, the ASR performance and the user-friendliness of the system are important factors. Furthermore, the acceptance of hearing loss by the hearing impaired person and the perceived seriousness of the hearing impairment influence the need for listening support (Southall, Gagné, & Leroux, 2006).

In Experiment 2, the ability to complete partly masked textual sentences (TRT) was related to lower task load when the captions were presented after the corresponding speech. Participants with good TRTs additionally reported that the text provided more speech comprehension support. Thus, the current results suggest that the ability to complete linguistic information is associated with lower subjective workload and more benefit obtained from the delayed captions. However, this relation was not observed in Experiment 1. We do not know what caused these different results; inspection of the correlation coefficients between the TRT and the mean NASA-TLX and text support ratings calculated separately for each of the lag-conditions indicated that the relation observed in Experiment 2 was based on both the low WA (60% and 70%) and the high WA (80% and 90%) conditions.

Note that compared with realistic online ASR, in Experiment 1, the text delay was already relatively short. The results of Experiment 2 suggest that this relatively short delay already increased the subjective task load considerably (Figure 4). These results indicate that when developing a communication system that automatically generates captions to support speech comprehension, the text delay should be minimized.

The test conditions included in the current study aimed to simulate daily telephone communication. Nevertheless, we used listen-only tests in our experiments. The main reason for doing this was that a Dutch ASR system yielding sufficient recognition performances in bidirectional communication was not available at the start of the current study. The speech comprehension benefit obtained from textual ASR output may be different in (simulated) bidirectional conversations as compared with the listen-only tests applied here. To be able to examine the benefit obtained from ASR output during conversations, the performance of ASR systems should increase. Training the ASR system with data (speech samples and correct transcriptions of those samples) that matches the speech in the tests will likely increase the recognition accuracy (Duchateau et al., 2005; Goronzy, 2002). Several differences between the listen-only laboratory tests applied in the current study and the use of ASR technology in real telephone conversations may affect the evaluation of the assistive technology by listeners with hearing impairment. For example, in bidirectional conversations, listeners have to attend to the speaking-turns in the conversation and they have to prepare their response to the incoming information, which may make it more difficult to attend to the textual information. On the other hand, several aspects of realistic bidirectional conversations may make speech comprehension less complex as compared to the listening tests applied in the current study. In normal conversations, the speaker is often familiar to the

listener and often, the listener knows the topic of the conversation at the start of the dialogue. These factors will increase speech comprehension, and may also improve the speech comprehension benefit obtained from the textual information. For example, knowledge about the speaker and the conversation topic could make it easier to ignore ASR errors that are clearly out of context. Another issue that should be examined in future studies is whether training the listener and the speaker to use the ASR application increases the benefit obtained from the text and improves the evaluation of the system. Several participants of the current study suggested that they were better able to ignore the ASR errors at the end of the test session.

Using "keyword spotting" instead of recognizing each spoken word may improve the speech comprehension benefit because keyword spotting results in less textual information that needs to be processed by the listener. In keyword spotting, the ASR system has a small vocabulary (typically around 100 words), and only this limited set of words can be recognized. However, out-of-vocabulary words do not result in recognition errors by default as compared with large-vocabulary ASR systems, such as the TNO ASR system used in the current study. This could also increase the (objective and/or subjective) speech comprehension benefit. Note that keyword spotting errors can occur, for example, when the system "misses" a keyword. A disadvantage of keyword spotting is that keywords have to be selected in advance, which can be difficult in actual conversations.

In conclusion, although our earlier studies showed that objectively measured sentence comprehension improves when partly incorrect captions are presented, the current findings indicate no subjectively evaluated speech comprehension benefit from textual ASR output when relatively long speech fragments are presented. For most conditions included in the current study, the task load neither decreased nor increased when listeners additionally had to process the visual information. The current study provides relevant insight in the key factors and difficulties when developing an assistive communication system based on ASR. Future research should aim at reducing the task load imposed by the task; otherwise, hearing impaired listeners may not be willing to use an assistive communication system providing erroneous textual information, despite any objective speech comprehension improvement. The present results underline that examining both objective and subjective speech comprehension benefit is required when developing systems intended to improve communication.

Acknowledgments

This study was supported by grants from the European Union FP6, Project 004171 HEARCOM. We thank J. H. M. van Beek for his technical assistance in the development of the tests.

Appendix A
NASA-TLX

In Table A.1, a description of the NASA-TLX and the ASR-output evaluation subscales is presented. Figure A.1 shows the mental demand subscale of the NASA-TLX.

Table A.1. NASA-Task Load Index Subscales

	Endpoints	Description
Subscale		
Mental demand	Low/high	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Time demand	Low/high	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Perfect/failure	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/high	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration level	Low/high	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?
ASR-output evaluation scales		
Text support	Support/hindrance	Did you obtain benefit from the text or did the text hinder speech comprehension?
Problems errors	Not problematic/severely problematic	To what extent were the ASR errors problematic for text comprehension?
Problems delay	Not problematic/severely problematic	To what extent was the text delay problematic for text comprehension?

NOTE: ASR = automatic speech recognition.

Figure A.1. The mental demand NASA-Task Load Index (NASA-TLX) subscale.

Appendix B
Results of the Questionnaire Presented at the End of the Test Session

The results on the Part 1 of the questionnaire concerning speech comprehension in daily telephone conversations are presented in Table B.1. The results of Experiments 1 and 2 are presented together.

As can be seen in Table B.1, participants reported fewer speech comprehension problems for telephone communication with hearing aids or assistive communication systems. They reported that speech comprehension decreases in the presence of background noise, and when the speaker is in background noise. Compared with telephone conversations by fixed phones, participants reported slightly less speech comprehension problems

(*continued*)

Appendix B (continued)

Table B.1. Results of Part 1 of the Questionnaire on Speech Comprehension Problems Experienced in Daily Life[a]

	No Speech Comprehension Problems	Speech Comprehension Problems in Background Noise	Speech Comprehension Problems in Quiet	I Am Not Able to Comprehend Speech; I Hear Sounds	
Q9 and Q14: Describe your hearing . . .					
Q9. . . . without hearing aids/assistive listening systems (n = 30)		33	43	23	
Q14. . . . with hearing aids/assistive listening systems (n = 20)	5	70	25		
	Excellent	Good	Moderate	Bad	Extremely Bad
Telephone conversation with hearing aids/assistive listening systems (n = 16)					
Q18. In silence	6	31	38	25	
Q19. In background noise		6	31	31	31
Q20. Speaker in background noise			44	50	6
Telephone conversation without hearing aids/ assistive listening systems (n = 30)					
Q21. In silence	3	30	40	20	7
Q22. In background noise		3	13	50	33
Q23. Speaker in background noise		10	27	50	13
Telephone conversation with mobile phone (n = 26)					
Q26. In silence	12	35	50	4	
Q27. In background noise	8	23	54	15	
Q28. Speaker in background noise		12	31	46	12

NOTE: Q = Question.

a. The results of Experiment 1 (n = 20) and Experiment 2 (n = 10) are presented together. All values are in percentages.

when using mobile phones. Half of the participants indicated that they communicate less often by a telephone due to their hearing impairment.

Part 2 of the questionnaire evaluated the captions and benefit obtained from the text. The majority (67%) of the participants indicated that the readability of the captions was high (Question [Q] 30). A total of 37% of the participants indicated that the appearance of the text on the screen "regularly" distracted them from listening to the speech, 30% answered "sometimes," and 20% answered "often" (Q31). About half of the participants preferred less words appearing on the screen simultaneously; for the others, the number of words appearing simultaneously was satisfactory (Q32). Most of the participants in Experiment 1 (65%) would not like to use the system at home; 30% doubted whether they would use it. In Experiment 2, 50% of the respondents reported not to like to use the system, but 30% would use the system at home (Q33), thus the willingness to use the system was higher for the participants in Experiment 2. The main reason for the reluctance to use the system was the number of ASR errors present in the text, followed by the text delay (Q33 and Q37). Positively evaluated system features were the presence of several text lines giving the participants more

(*continued*)

Appendix B (continued)

time to read the captions, and provision of several keywords by the text (Q36). The last question (Q44) announced a final Telephone-NC_{BB} test; participants were told that they would win a prize if they would answer three questions on the content of that fragment correctly. They were allowed to choose whether the captions were presented or not. In Experiment 1, 40% of the participants indicated that they would like the captions being presented, and in Experiment 2, this percentage equaled 80%. This again indicated that more participants of Experiment 2 were willing to use a system automatically providing captions during speech comprehension.

References

Altman, D. G. (1991). *Practical statistics for medical research.* London: Chapman & Hall.

Boothroyd, A. (2004). Hearing aid accessories for adults: The remote FM microphone. *Ear and Hearing, 25*, 22-33.

Bourke, P. A., Duncan, J., & Nimmo-Smith, I. (1996). A general factor involved in dual task performance decrement. *Quarterly Journal of Experimental Psychology Section A, 49*, 525-545.

Deaton, J. E., & Parasuraman, R. (1993). Sensory and cognitive vigilance: Effects of age on performance and subjective workload. *Human Performance, 6*, 71-79.

Duchateau, J., Van Uytsel, D. H., Van Hamme, H., & Wambacq, P. (2005, September). Statistical language models for large vocabulary spontaneous speech recognition in Dutch. In *Proceedings of INTERSPEECH-2005* (pp. 1301-1304), Lisbon, Portugal.

Fiscus, J., Garofolo, J., Lee, A., Martin, A., Pallett, M., Przybocki, M., et al. (2004, November). Results of the fall 2004 SST and MDE evaluation. *DARPA Rich Transcription Workshop*, Palisades, NY.

Fitzpatrick, R. (1991). Surveys of patient satisfaction: II—Designing a questionnaire and conducting a survey. *British Medical Journal, 302*, 1129-1132.

George, E. L. J., Zekveld, A. A., Kramer, S. E., Goverts, S. T., Festen, J. M., & Houtgast, T. (2007). Auditory and nonauditory factors affecting speech reception in noise by older listeners. *Journal of the Acoustical Society of America, 121*, 2362-2375.

Goronzy, S. (2002). *Robust adaptation to non-native accents in automatic speech recognition* (Lecture Notes on Artificial Intelligence, Vol. 2560). Berlin: Springer-Verlag.

Graham, R., & Carter, C. (2000). Comparison of speech input and manual control of in-car devices while on the move. *Personal Technologies, 4*, 155-164.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: Elsevier Science.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors,* 34, 429-439.

Jerger, J., Chmiel, R., Florin, E., Pirozzolo, F., & Wilson, N. (1996). Comparison of conventional amplification and an assistive listening device in elderly persons. *Ear and Hearing, 17*, 490-504.

Jobard, G., Vigneau, M., Mazoyer, B., & Tzourio-Mazoyer, N. (2007). Impact of modality and linguistic complexity during reading and listening tasks. *NeuroImage, 34*, 784-800.

Karis, D., & Dobroth, K. M. (1995). Psychological and human factors issues in the design of speech recognition systems. In A. Syrdal, R. Bennett, & S. Greenspan (Eds.), *Applied speech technology* (pp. 127-194). Boca Raton, FL: CRC Press.

Karlsson, I., Faulkner, A., & Salvi, G. (2003, September). SYNFACE: A talking face telephone. In *EUROSPEECH-2003* (pp. 1297-1300), Geneva, Switzerland.

Kepler, L. J., Terry, M., & Sweetman, R. H. (1992). Telephone usage in the hearing-impaired population. *Ear and Hearing, 13*, 311-319.

Kricos, P. B. (2006). Audiologic management of older adults with hearing loss and compromised cognitive/psychoacoustic auditory processing capabilities. *Trends in Amplification, 10*, 1-28.

Legatt, A. P., & Noyes, J. M. (2004). A holistic approach to the introduction of automatic speech recognition technology in ground combat vehicles. *Military Psychology, 16*, 81-97.

Leitch, D. (2008). *GIFT Atlantic Liberated Learning High School Project* (Final research report). Halifax, Nova Scotia, Canada: Saint Mary's University.

Leitch, D., & MacMillan, T. (2003). *Innovative technology and inclusion: Current issues and future directions for Liberated Learning Research* (Year IV research report on the Liberated Learning Initiative, Liberated Learning Project). Halifax, Nova Scotia, Canada: Saint Mary's University.

Levitt, H. (1994). Speech processing for physical and sensory disabilities. In D. B. Roe & J. G. Wilpon (Eds.), *Voice communication between humans and machines* (pp. 311-343). Washington, DC: National Academies Press.

Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics, 28*, 367-378.

Milchard, A. J., & Cullington, H. E. (2004). An investigation into the effect of limiting the frequency bandwidth of speech on speech recognition in adult cochlear implant users. *International Journal of Audiology, 43*, 356-362.

Möller, S. (2000). *Assessment and prediction of speech quality in telecommunications.* Boston: Kluwer Academic.

Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *International Journal of Aviation Psychology,* 5, 87-106.

Moroney, W. F., Biers, D. W., Eggemeier, F. T., & Mitchell, J. A. (1992). Comparison of two scoring procedures with the NASA Task Load index in a simulated flight task. In *Proceedings of the 1992 IEEE National Aerospace Electronics Conference* (pp. 734-740). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Murata, A. (1999). Identification of an acceptable mixture of key and speech inputs in bimodal interfaces. *International Journal of Human-Computer Interaction, 11*, 339-348.

Nusbaum, H. C., DeGroot, J., & Lee, L. (1995). Using speech recognition systems: Issues in cognitive engineering. In A. Syrdal, R. Bennett, & S. Greenspan (Eds.), *Applied speech technology* (pp. 127-194). Boca Raton, FL: CRC Press.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors,* 33, 17-33.

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia,* 28, 1021-1034.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America,* 97, 593-608.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18*, 43-52.

Saunders, G. H., Forsline, A., & Fausti S. A. (2004). Performance-perceptual test and its relationship to unaided reported handicap. *Ear and Hearing,* 25, 117-126.

Siciliano, C., Faulkner, A., & Williams, G. (2003). Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In *Proceedings of the Auditory-Visual Speech Processing Workshop* (pp. 205-208), St Jorioz, France.

Southall, K., Gagné, J., & Leroux, T. (2006). Factors that influence the use of assistance technologies by older adults who have a hearing loss. *International Journal of Audiology,* 45, 252-259.

Stouten, F., Duchteau, J., Martens, J.-P., & Wambacq, P. (2006). Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication, 48*, 1590-1606.

Svensson, E. (2001). Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine,* 33, 47-48.

Tomporowski, P. D. (2003). Performance and perceptions of workload among young and older adults: Effects of practice during cognitively demanding tasks. *Educational Gerontology,* 29, 447-466.

Van Boxtel, M. P. J., Van Beijsterveldt, C. E. M., Houx, P. J., Anteunis, L. C., Metsemakers, J. F. M., & Jolles, J. (2000). Mild hearing impairment can reduce verbal memory performance in a healthy adult population. *Journal of Clinical and Experimental Neuropsychology,* 22, 147-154.

Vandierendonck, A., Kemps, E., Fastame, M. C., & Szmalec, A. (2004). Working memory components of the Corsi blocks task. *British Journal of Psychology,* 95, 57-79.

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *Journal of the Acoustical Society of America,* 107, 1671-1684.

Wald, M. (2006). Captioning for Deaf and Hard of Hearing People by Editing Automatic Speech Recognition in Real Time, *Proceedings of 10th International Conference on Computers Helping People with Special Needs ICCHP 2006, LNCS 4061,* pp. 683-690.

Wickens, C. D. (1992). *Engineering psychology and human performance*. New York: HarperCollins.

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.

Zekveld, A. A., George, E. L. J., Kramer, S. E., Goverts, S. T., & Houtgast, T. (2007). The development of the Text Reception Threshold test: A visual analogue of the Speech Reception Threshold test. *Journal of Speech, Language, and Hearing Research,* 50, 576-584.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S. M. G., & Houtgast, T. (2008). The benefit obtained from visually displayed text from an automatic speech recognizer during listening to speech presented in noise. *Ear and Hearing,* 29, 838-852.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S. M. G., & Houtgast, T. (in press). The influence of age, hearing, and working memory on the speech comprehension benefit derived from an automatic speech recognizer. *Ear and Hearing*.

Zhang, Y., & Luximon, A. (2005). Subjective mental workload measures. *Ergonomia IJE&HF,* 27, 199-206.