

University of Groningen

Listening Effort

Pals, Carina

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Pals, C. (2016). *Listening Effort: The hidden costs and benefits of cochlear implants*. [Thesis fully internal (DIV), University of Groningen]. Rijksuniversiteit Groningen.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

CHAPTER 5

Effect of spectral resolution on speech intelligibility,
comprehension, and listening effort in cochlear-implant users

Carina Pals, Anastasios Sarampalis, Andy Beynon,

Thomas Stainsby, Deniz Başkent

ABSTRACT

Objectives. Previous research has shown speech intelligibility in cochlear implant (CI) users to improve with increasing spectral resolution up to 7 active electrodes, and plateau thereafter. Here we hypothesized that further increased spectral resolution may still further improve listening effort, even if intelligibility remains unchanged.

Design. Spectral resolution was manipulated by varying the number of active electrodes of the CI between 7 and 15. After a one-month familiarization period, the CI users performed two experiments. In Experiment 1, a dual-task paradigm was used to measure speech intelligibility and listening effort, reflected in the accuracy scores on the primary listening task and response times on the secondary visual task, respectively. In Experiment 2, a sentence verification task was used to measure speech comprehension and listening effort, reflected in accuracy scores and response times, respectively.

Results. In line with literature, speech intelligibility did not improve beyond 7 active electrodes. In contrast, speech comprehension, as reflected by the sentence verification task, improved up to 11 active electrodes. The dual-task measure of listening effort showed no improvement beyond 7 active electrodes, while the sentence verification task measure of listening effort revealed a systematic improvement for increased spectral resolution up to 11 active electrodes.

Conclusion. The sentence verification task results revealed a benefit of increased spectral resolution for both comprehension and listening effort, for conditions that typically show no improvement in speech intelligibility. This highlights both the potential benefits of improving spectral resolution for CI users, and the added value of clinical assessment tools that can reveal such benefits when traditional speech intelligibility measures show no improvement. The sentence verification task may be a good candidate for such a clinical tool.

Keywords: Cochlear implant, speech perception, listening effort, spectral resolution

INTRODUCTION

Everyday verbal communication requires the listener to perceive, comprehend, and reason about the message conveyed by the speaker before responding. Successful speech comprehension involves perceptual and cognitive processing, as well as the appropriate allocation of attentional resources and processing (effort), especially when the acoustic speech signal is compromised (Wingfield & Tun, 2007). In ideal listening conditions, speech is perceived clearly and comprehension is nearly effortless (Mattys et al., 2012; Wild et al., 2012). In non-ideal listening conditions, however, degradations of the speech signal limit the effectiveness of bottom-up perceptual processes, increasing reliance on top-down cognitive processes for compensation (e.g. Başkent, Clarke, et al., 2016; Broadbent, 1958; Downs & Crum, 1978; Rönneberg, 2003). Degraded speech perception can be facilitated by, for example; top-down repair mechanisms to restore interrupted speech (e.g. Bhargava, Gaudrain, & Başkent, 2014; Miller & Licklider, 1950; Samuel, 1981), the use of linguistic knowledge (e.g. Benard, Mensink, & Başkent, 2014; Hannemann, Obleser, & Eulitz, 2007), or situational or linguistic context (e.g. Dahan & Tanenhaus, 2004; Sheldon, Pichora-Fuller, & Schneider, 2008; Wingfield, Aberdeen, & Stine, 1991). While the recruitment of higher-order cognitive processes can aid, and thus enhance, the comprehension of degraded speech, it may come at the cost of increased cognitive load (e.g. Hornsby, 2013; Pals, Sarampalis, & Başkent, 2013; Wingfield et al., 2007; Winn, Edwards, & Litovsky, 2015; Zekveld, Kramer, & Festen, 2010). This may in turn reduce the cognitive resources available for concurrent tasks (Sarampalis et al., 2009), lead to fatigue (Hornsby, 2013), affect the ability to remember the speech (McCoy et al., 2005; Rabbitt, 1966), and lead to slower speech comprehension (Mattys & Wiget, 2011; Wagner et al., 2016).

For cochlear implant (CI) users, signal degradation is an everyday occurrence. The quality of the CI-transmitted speech signal is affected by many factors, including, but not limited to, electrode placement, auditory nerve survival, as well as device-related factors such as front-end processing or electrode design (e.g. Başkent et al., 2016; Blamey et al., 1992). One of the most notable consequences is a severe reduction in spectral resolution as channel interactions limit the effective number of spectral channels (Stickney et al., 2006). The effect of spectral resolution on speech intelligibility has been studied extensively over the decades since the introduction of multichannel CIs (e.g. Eddington, 1980; Fishman, Shannon, & Slattery, 1997;

Friesen et al., 2001; Fu, Shannon, & Wang, 1998; Schwartz, Chatterjee, & Gordon-Salant, 2008; Winn, Chatterjee, & Idsardi, 2012). Research has shown, for example, that the recognition of individual phonemes in quiet improves up to 7 electrodes (Fishman et al., 1997) and thresholds for phoneme recognition in noise continue to improve up to, and possibly beyond, 16 electrodes (Fu et al., 1998). Sentence intelligibility, on the other hand, reaches ceiling performance at about 10 active electrodes for speech in noise (Friesen et al., 2001). While in quiet, sentence intelligibility plateaus around 4 active electrodes for the average CI users (Fishman et al., 1997), and around 7 active electrodes for high-performing CI users (Friesen et al., 2001). This increased tolerance for reduced spectral resolution when listening to full sentences can be explained by the availability of sentence context and suggests the involvement of effortful top-down processing to enhance intelligibility (Sheldon et al., 2008). This implies that while intelligibility has reached ceiling, listening effort may still be high. Effects of spectral resolution on listening effort for CI users, however, have not been previously documented.

The current study aims to investigate the effect of spectral resolution in CI users on not only speech understanding, but also listening effort. Specifically, we hypothesized that when spectral resolution increases, this could provide a benefit in listening effort, even after intelligibility performance has reached a plateau. Indirectly supporting this idea is the fact that in normal hearing (NH) listeners presented with acoustic, noise-band vocoder simulations of CI speech in quiet, speech intelligibility improved up to 6 spectral channels and plateaued thereafter, while listening effort continued to improve up to 8 spectral channels (Pals et al., 2013, see also Chapter 2). Similar results have been shown using pupil dilation as a measure of effort: in NH listeners, spectral resolution affects pupil dilation, i.e. listening effort, even when intelligibility is at 100% (Winn et al., 2015). In the present study, we systematically investigate whether, similar to NH listeners, CI users benefit from increased spectral resolution in reduced listening effort even when changes in intelligibility are not observed or expected.

In the present study, a dual-task paradigm first designed and used in our earlier study in NH listeners (Pals et al., 2013), was employed to measure intelligibility (primary listening task) and listening effort (secondary task) simultaneously. The dual-task paradigm is an established method for measuring cognitive load, and has been used to quantify listening effort in a

number of studies (e.g. Anderson Gosselin & Gagné, 2010; Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2010; Pals et al., 2013; Rakerd, Seitz, & Whearty, 1996; Sarampalis et al., 2009). The assumption that cognitive resources are limited and shared across tasks (Broadbent, 1958; Kahneman, 1973), implies that when two tasks are performed simultaneously, the execution of the primary task uses resources that would otherwise have been available for the secondary task. Performance on the secondary task thus reflects the cognitive processing load of the primary task (Broadbent, 1958; Rabbitt, 1966). The current dual-task paradigm was successfully used by Pals et al. (2013) in support of the present hypothesis using acoustic simulations in a homogenous group of young adult NH listeners. The question remains whether the method is suitable for use with CI users, since a range of different factors can affect performance in CI users (Başkent, Gaudrain, et al., 2016), and effects of age may further affect the results as CI users tend to be older (Bhargava et al., 2014; Bhargava, Gaudrain, & Başkent, 2016).

Therefore, the dual-task paradigm (Experiment 1) was complemented with a simpler, single-task measure of comprehension and processing speed (Experiment 2); the sentence verification task (Adank & Janse, 2009; Baer, Moore, & Gatehouse, 1993). While this task was not previously used with CI users, a version of this task has successfully been applied in previous research to reveal effects of hearing-aid processing on listening effort in elderly (age 60+) hearing impaired participants (Baer et al., 1993). In the sentence verification task, participants listen to sentences that are either unmistakably true or false/nonsense. The task requires the listener to respond via key-press indicating whether the sentence they heard was true or false/nonsense, producing both accuracy scores and response times (RTs). As an increase in cognitive load leads to slower comprehension (Gibbon, Moore, & Winski, 1997; Mattys et al., 2011; Wagner et al., 2016), the sentence verification accuracy and RTs can be interpreted to reflect comprehension and cognitive processing load, i.e. listening effort, respectively.

Overall, we hypothesize that reduced spectral resolution in CI users will have a detrimental effect, not only on speech understanding, but also on listening effort. Crucially, similar to the findings in NH listeners (Pals et al., 2013), we expect that listening effort could be improved further with increasing spectral resolution even when intelligibility appears unchanged.

EXPERIMENT 1: DUAL-TASK APPROACH: SPEECH INTELLIGIBILITY
AND LISTENING EFFORT

To be able to compare our results with NH listeners and CI users, the same dual-task paradigm we had designed for our previous study (Pals et al., 2013) was used. A few minor modifications were made to the design to accommodate for expected differences in speech understanding and response speed between the young NH participants of the previous study and the adult and elderly CI user participants of this study. Specifically, easier sentence materials were used and the response time-out was longer; these changes are described in more detail below.

Methods

Participants. Initially, a total of 34 CI users were recruited for participation, 17 through the Audiology Department at the University Medical Center Groningen and 17 through the Audiology Department at the Radboud University Medical Center in Nijmegen. Of the participants recruited in Groningen, three served as pilot participants, two could not come back for the second session due to health reasons, two could not complete the experiment due to a technical problem, and one was unable to follow the test instructions. The data from the remaining 9 participants were included in the final analyses. From the participants recruited in Nijmegen one did not return for the second session and the data from the remaining 16 were included in the final analysis. This resulted in a total of 25 participants (14 female, mean age 58 years, range 34 - 76) who completed the two experiments fully without any problems.

The participants were all native Dutch speakers, and postlingually deafened adults, implanted with the Cochlear Nucleus device, and using the CP810 processor. Two participants had been hearing impaired since birth (marked by asterisk in the Table 1), however, all learned their native language in audio-verbal mode. As the goal of this study was to investigate listening effort for speech understanding at ceiling, only the best performing CI users were chosen. Inclusion criteria were a minimum of 1 year experience with CI use, clinical consonant-nucleus-consonant word recognition scores of 80% or higher, and no known cognitive disabilities. All participants had either normal vision, or vision corrected to normal by spectacles. All-but-one of the participants had complete intra-cochlear electrode array insertion and all were fitted with at least 15 active electrodes in their daily speech processor

maps. This and the subsequent experiments were both approved by the local ethical committee. Demographic and hearing-related information for these participants is summarized in Table 1.

Table 1. Summary of the CI participants' demographic and hearing-related information.

Subject ID	Gender	Age during experiment	Age of first HL	Duration of CI use	Etiology
304	M	38	3	2.3	Usher
307	M	64	46	1	Progressive
310**	F	54	49	5	Wegener
311	M	59	31	2	Meningitis
313*	M	60	0	7	Mother rubella
314	M	51	7	12	Osteoporosis
315	F	69	33	7	Progressive
316	F	41	6	2	Hereditary
317	M	76	10	2	Otitis media
321	F	51	10	8	Progressive
322	F	59	54	2	Schwannoma
323	F	67	38	7	Stapedectomy
324	F	66	38	3	Progressive
325	F	52	26	2	Progressive
326	M	62	38	3	Progressive
327	F	70	14	4	Progressive
328	M	34	65	4	Progressive
329	M	65	16	7	Progressive
330	F	58	48	6	Progressive
331	M	67	43	3	Progressive
332	F	59	58	7	Progressive
333	M	65	40	4	Progressive
334	F	58	34	4	Otosclerosis
335*	F	49	0	17	Hereditary
336	F	62	30	3	Ototoxicity

Note: * CI users who were hearing impaired since birth. ** CI user who did not have a fully inserted electrode array.

Speech stimuli. In our previous study with normal-hearing (NH) participants we used sentences from the VU corpus (Vrije Universiteit; Versfeld et al., 2000). However, even CI users selected for high phoneme scores may still show poor sentence intelligibility for the VU corpus speech materials (Bhargava et al., 2014, 2016). In the current study, the speech stimuli for the primary intelligibility task were therefore taken from the LIST corpus (van Wieringen & Wouters, 2008). This corpus is specifically optimized to provide accurate speech reception thresholds for Dutch and Flemish hearing-impaired listeners and CI users in quiet and in noise. The corpus consists of 35 lists of 10 everyday conversational Dutch sentences, each spoken by the same female speaker. The lists are balanced for equal difficulty. The total

number of syllables in each list of 10 sentences is 90. The lists are structured such that the first sentence is short (between 4 and 6 syllables) and each consecutive sentence is one or two syllables longer than the previous one, ending with a long sentence (between 12 and 15 syllables).

Visual stimuli. The visual stimuli for the secondary rhyme-judgment task were monosyllabic Dutch words. The lists of words used in this experiment were compiled by Pals et. al. (2013), and consist of rhyme words for several word endings for each of the 5 basic Dutch vowels (a, e, i, u, o). Each word list was examined by a native Dutch speaker, and words with multiple possible pronunciations, as well as the 25 least common words according to the CELEX lexical database of Dutch (Baayen, Piepenbrock, & van Rijn, 1993) were excluded (Pals et al., 2013). In the experiment, the words were presented one above the other in black capital letters on a white background on a computer monitor approximately 50cm in front of the participant. The letters were approximately 9mm high and 7mm wide, with 12 mm whitespace between the two words.

Stimulus presentation and equipment. The experiment was programmed in MATLAB using Psychtoolbox Version 3, and ran on a Macbook Pro 2010 laptop. The program coordinated the presentation of the speech and visual stimuli and logged the responses and response times on the secondary task. The verbal responses on the primary speech task were recorded using a digital audio recorder to be scored later by a native Dutch speaker. The experiment was conducted in a sound-isolated booth. All speech stimuli were presented directly from the experimental computer via personal audio cable to the CI processor, to avoid small differences in residual hearing affecting the outcome. As a result, the stimulus presentation level was not controlled for in dB SPL but set to a comfortably loud level at the start of the experiment and kept the same throughout data collection.

Experimental conditions. Spectral resolution was manipulated by altering the number of active electrodes of the CI by disabling electrodes and redistributing the frequencies assigned to them to the remaining electrodes (Friesen et al., 2001). Previous research has shown that, on average, CI users' speech intelligibility performance in quiet is near ceiling from about 7 active electrodes (Fishman et al., 1997; Friesen et al., 2001). Because a core question of this study is whether changes in listening effort occur when intelligibility no longer improves, the

experimental conditions were therefore chosen to cover the range between 7 electrodes and the full 22-electrode array. Specifically, four experimental maps were generated with 7, 9, 11, and 15 active electrodes, chosen because these numbers allowed for the active electrodes to be either evenly spaced or distributed in a regularly recurring pattern across the full 22-electrode array (Figure 1). The experimental maps were generated based on the participant’s own preferred map using Cochlear Corp’s Custom Sound software (version 4.0), and the frequencies were redistributed over the active electrodes as suggested by the software, which resulted in a redistribution of frequencies similar to that used by Friesen et al. (2001). All other parameters (T and C values, stimulation rate, pulse width, coding strategy) were left unchanged. The participant’s preferred SmartSound features, such as noise reduction, AutoSens, ADRO, etc., were also left as is.

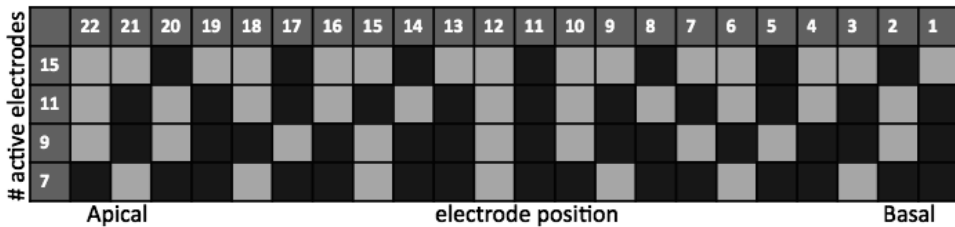


Figure 1. The distribution of active electrodes along the full array is shown for each of the experimental conditions. A light gray square denotes an active electrode, a dark gray square a deactivated electrode.

Procedure. The experiment consisted of two testing sessions in which the participants performed both Experiment 1 and 2 (Experiment 2 will be described later), with a one-month training period in-between. During the one-month training period between the two sessions, the participants received the experimental processor with the four experimental maps to take home. They were instructed to practice listening with the maps for one hour a day, rotating between the four different maps on the processor. This served to familiarize the listener with the experimental maps before the actual testing session, thus minimizing acute effects of new, unfamiliar stimulation patterns and training effects over the course of the experiment. Research shows that, in the case of spectral mismatch, familiarization occurs relatively fast over the first few days or weeks when the experimental processor is used all day long (Fu, Shannon, & Galvin, 2002). As the reduced spectral resolution of our experimental programs may negatively impact the CI participants’ listening abilities for example at the workplace, we decided instead to limit familiarization to one hour a day, for one month. To verify whether

the participants had been practicing with the experimental processor, they were asked a few questions at the start of the second session. The participants were asked about their experiences with the experimental processor, whether they had experienced any difficulties, and whether they had noticed distinct differences between the programs.

The first session lasted at most 1 hour, during which the participants were tested using their preferred map on their own processor, while simultaneously the experimental processor was programmed. The second session lasted approximately 2 hours, during which the participants were tested with each of the 4 experimental maps, in counterbalanced order (in a 4x4 balanced Latin-square design).

At the start of the first session, after explaining the procedure and allowing for questions, the presentation level for the speech stimuli was determined. A sample sentence was played repeatedly, starting at a very low presentation level and increasing in steps of 2.5 dB. Following clinical procedure, each time the sentence was presented, the participants were asked to indicate the perceived loudness on a visual scale ranging from “imperceptibly soft” to “uncomfortably loud”. When a comfortably loud level was reached, the stimulus was presented another three or four times, alternately increasing and decreasing in loudness by 2.5 dB to confirm that the selected level was loud and clear, yet still comfortable. After this, while the participants performed the experimental tasks with their own processor using their preferred map, the experimental processor was programmed based on this preferred map.

At the start of each session, the procedures of the two tasks were explained and participants performed a 3-minute training session for the rhyme-judgment task before starting the actual experiment. Each condition was tested in a series of four task blocks. First, the intelligibility task was presented twice alone (single task), one training block and one experimental block, then the intelligibility task and secondary rhyme-judgment task were presented twice simultaneously (dual task), first a training block and then an experimental block. For each of the experimental conditions, the participants completed the full series of 4 task blocks before moving on to the next condition.

The primary intelligibility task required the participants to listen to the sentence stimuli and repeat them out loud, giving their best guess when they were not sure what they heard. When

the intelligibility task was presented alone, one list of 10 sentences was used. When presented simultaneously with the secondary task, one list of 10 sentences was used for training and two lists of 10 sentences each were used for the experiment. The sentences varied considerably in duration, unlike the sentences used by Pals et al. (2013), and therefore needed a different strategy for silent interval duration than the study by Pals et al. (2013). The sentences in this study were followed by a silent interval of the duration of the sentence recording plus an additional 2.5 seconds. This provided the participants sufficient time to repeat the sentence before the next sentence was presented.

In the secondary visual rhyme-judgment task, a pair of words was presented on the screen. The task was to answer as fast as possible whether the word pair rhymed or not, by pressing either 'v' for yes or 'n' for no on a keyboard. These keys were chosen for their convenient position at the front edge of the keyboard. The word pair was randomly chosen by the MATLAB program, with a 50% chance of a rhyming pair. The stimuli were presented until a key was pressed, or until the time-out of 5 seconds was reached. The time-out was longer than in our previous study to accommodate the more advanced age of some of the participants of the present study. If after these 5 seconds no key was pressed, this was logged as 'unanswered'. After each stimulus, a fixation cross was presented on the screen for a random duration between 0.5 and 2.0 seconds before moving on to the next word pair.

In the dual-task, the participants were instructed to perform the listening task and the rhyme-judgment task simultaneously. Following the design of the previous study, participants were instructed to prioritize the primary listening task over the secondary rhyming task and to respond to the secondary task as fast as possible. Because of the independent timing of the two tasks, secondary rhyme-judgment task trials could occur both during and between the presentations of sentences.

Results

The left panel of Figure 2 shows the intelligibility scores for the primary listening task, both in single task (open symbols) and in dual task (filled symbols). The baseline included in the graph reflects the average intelligibility score when the CI users were tested with their own preferred map using the full electrode array. Because the baseline scores were recorded in the first session of the experiment, and not as part of the actual data collection (i.e., within the counter-

balanced test conditions), these were not included as a condition in the analysis. They are shown here purely as a reference level. The speech intelligibility scores from experimental conditions were analyzed using a two-way repeated-measures ANOVA using R and the ez package (version 4.2-2) including the main factors spectral resolution (4 levels: 7, 9, 11, 15 active electrodes) and task type (2 levels: single or dual task), and presentation order as a covariate. The ANOVA revealed no significant effects of spectral resolution or task type on speech intelligibility and no significant interaction.

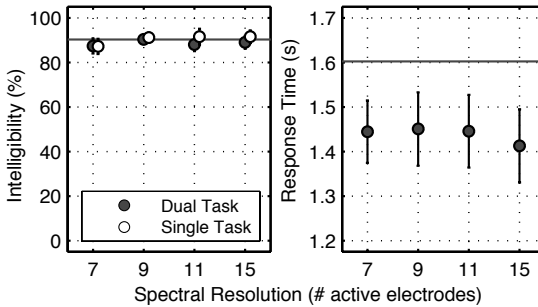


Figure 2: The left panel shows the speech intelligibility in percentage sentences correctly repeated, for both single task (open symbols) and dual task (filled symbols), as a function of spectral resolution. The right panel shows the response times in seconds on the dual-task secondary task. Error bars in both panels denote standard errors. The lines show the average baseline performances for the participants when tested with their own device in the first session of the study.

The right panel of Figure 2 shows the RTs on the secondary rhyme-judgment task in the dual-task. For the RTs on the secondary rhyme-judgment task, the number of observations per participant per condition varied depending on the response speed and accuracy. The analysis method of choice for data with different number of observations per cell is linear mixed effect (LME) models. The RTs were analyzed using R and the lme4 package (version 1.1-7, lmerTest-package version 2.0-11). To approximate a normal distribution, the data were log-transformed by taking the natural logarithm of the RTs. The log-transformed RTs (lnRTs) approximated a normal distribution for RTs between 0.35 and 3 s but deviated from normal outside that range, therefore RTs below 0.35 and over 3 s were excluded (5.9% of all trials). Accuracy on the rhyme-judgment task varied slightly, between 94% and 96%, and only trials with correct responses were included in the analysis of RTs. However, to account for differences in accuracy between participants and conditions, the accuracy scores were

included as a factor in the model. As with the speech intelligibility scores, the baseline RTs recorded in the first session were not included as a condition in the analysis, however, when included as a factor in the model they contributed significantly to the fit of the model and were therefore included ($\chi^2(1) = 36.202, p < 0.001$).

The final model included the factors spectral resolution, presentation order, accuracy, and baseline RT. A random intercept was included for participantID, and random slopes and intercepts were included for all within-subject factors. The intercept of the model corresponds to average difference in RT compared to baseline on the secondary task while listening to speech using 7 active electrodes as the first task of the experiment, and did not differ significantly from 0 ($\beta = -0.1194, SE = 0.0769, t = 1.554, p = 1.256$). The model revealed a significant effect of presentation order on lnRT, estimated at 0.0182 ($\beta = 0.0182, SE = 0.0083, t = -2.208, p = 0.038$), suggesting a decrease in RTs of $e^{(-0.1194-0.0182)} \cdot e^{-0.1194} = -0.160$ s for each consecutive condition in the experiment, and a significant effect of baseline RT, estimated at 0.2487 ($\beta = 0.2487, SE = 0.0297, t = 8.375, p < 0.001$), suggesting that participants with higher baseline RTs also have longer RTs in the experiment overall ($e^{(-0.1194+0.2487)} \cdot e^{-0.1194} = 0.2506$ s longer RTs in the experiment per 1 second longer baseline RTs). The model showed no significant effect of spectral resolution ($\beta = -0.0037, SE = 0.0022, t = -1.670, p = 0.109$), or accuracy ($\beta = -0.0062, SE = 0.0064, t = -0.971, p = 0.336$) on RT.

EXPERIMENT 2: SVT APPROACH: SPEECH COMPREHENSION AND LISTENING EFFORT

In Experiment 1, we had used the dual-task paradigm, as it had been previously tested and validated with NH participants listening to CI simulated speech (Pals et al., 2013). The sentence verification task we used in Experiment 2 had not been used with CI-simulated speech before. Therefore, an additional group of NH participants was recruited for Experiment 2 only, to evaluate this specific task as a measure of listening effort in NH listeners and to examine how it reflects the effects of reduced spectral resolution in NH listeners presented with CI-simulated speech.

Methods

Participants. Experiment 2 was performed by two groups of participants: a group of 24 young adult NH listeners and the same 25 CI users that participated in Experiment 1.

Initially, 25 NH listeners were recruited for this experiment, all students of the Psychology Department of the University of Groningen, and they received partial course credit for their participation. One of the participants was excluded because of missing data due to a technical error during the experiment. The remaining 24 participants (4 male) were all native Dutch speakers and young adults (mean age 21 years, range 19 – 27). All NH participants had hearing thresholds of 20 dB HL or better at all audiometric frequencies between 250 and 6000 Hz. Exclusion criteria were self-reported dyslexia and other language disabilities.

Speech stimuli. The sentence material used for the sentence verification task was created by Adank and Janse (2009), and the same recordings were used for both the NH and the CI participants. The corpus consists of in total 180 sentences, all spoken at a normal conversational speaking rate by the same male native Dutch speaker. The sentences are all syntactically correct, however, 90 are unarguably true and make sense (e.g. *Tijgers hebben een staart*, Tigers have a tail), and the other 90 are obviously false or nonsense (e.g. *Een aap is een soort vis*, A monkey is a type of fish). All sentences start with the subject noun followed by a predicate, are at least 3 words long (min. 4 syllables), and the longest sentence is 8 words long (max. 14 syllables).

Stimulus presentation and equipment. The experiment was programmed, presented, and logged in the same manner as Experiment 1. For the NH participants, the speech stimuli were presented via an AudioFire 4 external soundcard of Echo Digital Audio Corporation (California, USA) and a DA10 digital-to-analog converter of Lavry Engineering, Inc. (Washington, USA) to the open-back HD600 headphones of Sennheiser electronic GmbH & Co. KG (Wedemark, Germany) at 65 dB A.

For the CI users, stimuli were presented in the same way and at the same level as for Experiment 1.

Experimental conditions. For the NH listeners, spectral resolution was manipulated by varying the number of bands of noise-vocoded CI simulation. The auditory stimuli were presented in 6 conditions; 4-, 6-, 8-, 12-, and 16-band noise-vocoded CI-simulated speech, and an unprocessed baseline condition, this was a subset of the same conditions used in our previous dual-task study (Pals et al., 2013). All speech stimuli, including the unprocessed condition, were band-pass filtered to 80 – 6000 Hz. The CI simulations were generated using the method as described by Shannon et al. (1995). For each of the CI simulation conditions, the 80 – 6000 Hz frequency range was divided into the desired number of bands such that the bands, from lower to upper -3 dB cut-off frequency, spanned approximately equal distances in the cochlea according to the Greenwood function (Greenwood, 1990). The speech recording was band-pass filtered into the desired number of analysis bands, using 6th order Butterworth band-pass filters. The noise carriers were generated by filtering white noise into bands using the same band-pass filters. From each of the analysis bands the envelope was extracted using half-wave rectification and low-pass filtering at 160 Hz using a 3rd order Butterworth filter. The carrier noise bands were modulated using the envelopes of the corresponding analysis bands and post-filtering using the original band-pass filters, and finally the resulting bands were combined to form the noise band vocoded CI simulation speech signal.

For the CI users, the experimental conditions of varying spectral resolution were the same as in Experiment 1, described above.

Procedure. All NH and CI participants were tested with a similar procedure. They were instructed to listen to one sentence at a time, and to indicate whether the sentence was true or false/nonsense by pressing either ‘v’ for true or ‘n’ for false/nonsense. The participants were instructed to respond as accurately and fast as possible. Whether a true or false sentence was played was determined randomly by MATLAB, with a 50% chance for either. The experimental program logged the responses and recorded the RTs from the end of the stimulus to the button-press, following previous research using the same paradigm (Adank et al., 2009), this implies that RTs were possible. If no key was pressed 5 seconds after start of the sentence, the program logged this as a ‘miss’ and moved on to the next sentence. A silent interval of random duration between 1.5 and 3.0 seconds was used between the end of the trial and the presentation of the next sentence stimulus.

The NH participants performed Experiment 2 in one session, which lasted approximately 1 hour. The CI users performed Experiment 2 in two sessions, with a one-month training period in-between, similar to Experiment 1. Session one lasted about 1 hour, and session two about 2 hours. They performed Experiment 1 and 2 one after the other in session 1 with their own processor, and after the training period in session 2 with the experimental maps on the experimental processor in an interleaved fashion; for each of the 4 experimental maps, the tasks for both Experiment 1 and 2 were performed before moving on to the next map. To minimize any effects of condition order, one half of the participants performed the dual task first, followed by the sentence verification task, and the other half did the opposite. At the start of each session, the task was explained verbally, followed by one training block consisting of 15 sentences for the first session and 10 sentences for the second session. The experimental blocks were presented in counterbalanced order and consisted of 30 sentences each, of which the first 5 sentences were considered training and were not included in the performance score of the task, resulting in 25 sentences per condition.

Results

NH listeners. The accuracy data were converted to the sensitivity measure d' , because this provides a bias-free measure of accuracy, i.e. it is not affected by individual preferences for either 'yes' or 'no' answers that may distort a % correct accuracy score. Figure 2, top-left panel shows the accuracy in d' scores for the sentence verification task for the NH listeners. The baseline included in the graph reflects the average accuracy using unprocessed speech stimuli. A one-way repeated-measures ANOVA with spectral resolution (4-, 6-, 8-, 12-, and 16-band noise vocoder CI simulation) as a numerical within-subject factor and covariate task order, revealed a significant effect of spectral resolution ($F(1, 23) = 36.696, p < 0.001$).

In order to examine the relationship between spectral resolution and accuracy, the results were modeled using a linear model including the within-subject factors spectral resolution (4, 6, 8, 12, 16 channel CI simulations) and task order, and a random intercept for participant ID as well as a random slope for spectral resolution per participant ID. Including baseline score did not contribute to the fit of the model ($\chi^2(1) = 0.0151, p = 0.9021$) and was therefore, for the sake of simplicity, not included in the final model.

The final model's intercept, corresponding to the average accuracy (in d') for 4 channel conditions, was estimated at approximately 2.86 ($\beta = 2.8564$, $SE = 0.2844$, $t = 10.044$, $p < 0.001$) and the effect of number of channels at 0.20 ($\beta = 0.2021$, $SE = 0.0259$, $t = 7.807$, $p < 0.001$), suggesting a 0.20 d' increase in accuracy for every additional channel in the CI simulation. No significant effect of task order was found ($\beta = 0.0778$, $SE = 0.0647$, $t = -1.203$, $p = 0.232$).

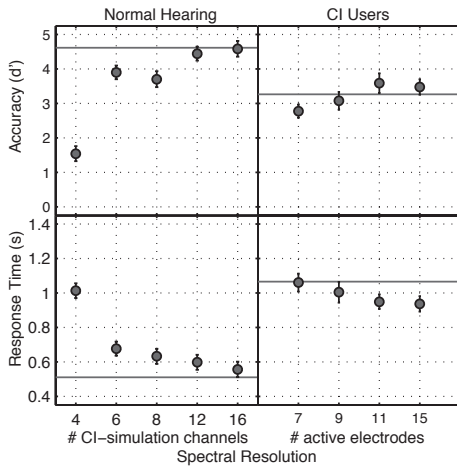


Figure 3: Results of the sentence verification task shown for NH participants (left-side panels) and CI participants (right-side panels). The top panels show accuracy scores in d' and the lower panels show RTs. Error bars show standard error. The baselines included in each figure show the average score for unprocessed speech for NH participants, and for the CI users the average score when tested with their own device.

The lower-left panel of Figure 3 shows the RTs on the sentence verification task for the NH listeners. The RTs approximated a normal distribution between -0.1 and 2.15 seconds, deviating from normal outside that range. Therefore, RTs under -0.1 and above 2.15 s were excluded from the analysis. This amounted to 2.7% of the responses. Because only correct responses were included and the longer and very short RTs were excluded, the number of observations varied per participant per condition. The RT data were therefore analyzed using LME models. The best fitting model for the RTs included the factors spectral resolution, presentation order, and baseline RT, as well as random intercepts for participant ID and sentence ID, and random slopes for spectral resolution for both participant ID and sentence ID.

The model's intercept was estimated at 512 ms ($\beta = 0.5123$, $SE = 0.0574$, $t = 8.926$, $p < 0.001$) and corresponds to the estimated average difference in RTs compared to baseline for the 4-channel CI simulation when presented as the first task of the experiment. The model showed a significant effect of spectral resolution, estimated at -25 ms ($\beta = -0.0252$, $SE = 0.0031$, $t = -8.073$, $p < 0.001$) suggesting a 25 ms decrease in RT for each additional spectral channel. The model also revealed a significant effect of baseline RT, estimated at 569 ms ($\beta = 0.5693$, $SE = 0.0892$, $t = 6.381$, $p < 0.001$), suggesting that participants with longer baseline RTs responded more slowly during the experiment as well (1 s longer baseline RT predicts on average 569 ms longer RTs in the experiment). The effect of presentation order was not significant ($\beta = -0.0025$, $SE = 0.0041$, $t = -0.604$, $p = 0.546$).

Because the relationship between the spectral resolution of the CI simulation and RT on the sentence verification task appears to be linear from 6 spectral channels up, but with a sharp increase in RTs from 6 to 4 channels, the results were re-modeled excluding the 4 channel condition, in order to see whether the effect would still be significant. The new model's intercept was estimated at 367 ms ($\beta = 0.3672$, $SE = 0.0570$, $t = 6.444$, $p < 0.001$) and corresponds to the estimated average difference in RTs compared to baseline for the 6-channel CI simulation when presented as the first task of the experiment. The model showed a significant effect of number of channels, estimated at -12 ms ($\beta = -0.0118$, $SE = 0.0023$, $t = -5.166$, $p < 0.001$), and a significant effect of baseline RT, estimated at 603 ms ($\beta = 0.6032$, $SE = 0.0922$, $t = 6.537$, $p < 0.001$). The effect of presentation order was again not significant ($\beta = -0.0062$, $SE = 0.0041$, $t = -1.509$, $p = 0.132$).

CI users. The top-right panel of Figure 3 shows the accuracy in the sentence verification task with CI users in d' scores. The baseline reflects the average accuracy recorded in the first session with the full electrode array. A one-way repeated-measures ANOVA with numerical within-subject factor spectral resolution and covariate task order showed a significant effect of spectral resolution on accuracy ($F(1, 24) = 15.510$, $p < 0.001$). In order to examine the effect of spectral resolution on accuracy, the results were modeled using a linear model. Including baseline RT as a factor did not improve the model fit ($\chi^2(1) = 3.7594$, $p = 0.053$) and was therefore not included in the model.

The final model, with within-subject factors spectral resolution (7, 9, 11, 15 active electrodes) and task order, a random intercept for participant ID as well as random slope for spectral resolution per participant ID, estimated the intercept at an accuracy score of $d' = 2.297$ ($\beta = 2.2967$, $SE = 0.3986$, $t = 5.762$, $p < 0.001$), corresponding with the estimated accuracy for 7 active electrodes when presented as the first task of the session. The model showed a significant effect of spectral resolution on accuracy of 0.129 ($\beta = 0.1291$, $SE = 0.0405$, $t = 3.187$, $p = 0.002$), suggesting an increase in d' of 0.129 for each additional active electrode. The effect of task order was not significant ($\beta = 0.1291$, $SE = 0.0405$, $t = 3.187$, $p = 0.072$).

The lower-right panel of Figure 3 shows the RTs in the sentence verification task with CI users, with the average RT recorded in the first session, with the full electrode array, included as a baseline. Only RTs for correct trials were included in the analysis. The RTs approximated a normal distribution between -0.2 and 3.2 seconds. RTs outside the range -0.2 and 3.2 s deviated from the normal distribution and were therefore excluded from the analysis. This amounted to 0.5% of the responses. The best fitting LME model for the RTs included the factors spectral resolution, presentation order, and baseline RT, as well as random intercepts for participant ID and sentence ID, and random slopes for spectral resolution for both participant ID and sentence ID.

The model's intercept was estimated at 774 ms ($\beta = 0.7741$, $SE = 0.1298$, $t = 5.964$, $p < 0.001$), and corresponds with the estimated difference in RT compared to baseline for the 7 active electrodes condition when presented as the first task of the experiment. The effect of number of channels was estimated at -17 ms ($\beta = -0.0169$, $SE = 0.0058$, $t = -2.884$, $p = 0.007$) suggesting a 17 ms decrease in RTs for each additional active electrode. The effect of presentation order was estimated at -58 ms ($\beta = -0.0582$, $SE = 0.0100$, $t = -5.800$, $p < 0.001$), suggesting a 58 ms decrease in RTs for each consecutive block in the experiment. The effect of baseline RT was estimated at 407 ms ($\beta = 0.4074$, $SE = 0.1003$, $t = 4.062$, $p < 0.001$).

DISCUSSION

The current study investigated how spectral resolution affects speech intelligibility, speech comprehension, and listening effort in CI users. In our previous study in NH listeners, we observed that, even when intelligibility had already reached ceiling, further improved spectral

resolution could still further improve listening effort. Based on this observation, we hypothesized that for CI users, listening effort may similarly improve with increased spectral resolution, even when speech intelligibility is near ceiling. Experiment 1 examined the effect of spectral resolution on speech intelligibility and listening effort in CI users, using a dual-task paradigm that was validated in our previous study with NH participants listening to noise-vocoded CI simulated speech (Pals et al., 2013). However, this dual-task paradigm had not been used with CI users before. The differences between the NH participants of our previous study and the CI participants of this study, both in hearing ability and age, might affect the dual-task outcome, possibly resulting in floor or ceiling effects. We therefore included a second experiment, using a simple, single-task, response-time measure of listening effort. Experiment 2 examined the effects of reduced spectral resolution on speech comprehension and processing speed in both NH and CI listeners, using a sentence verification task. The results in a nutshell: Experiment 1 showed no effect of spectral resolution on either intelligibility or listening effort; Experiment 2, on the other hand, showed a clear effect of spectral resolution on both speech comprehension and processing speed in NH as well as CI participants. Each of these findings will be discussed in more detail below.

The results from Experiment 1 showed that for CI users, further increased spectral resolution from 7 active electrodes upwards, did not lead to further improved sentence intelligibility in quiet listening conditions. These findings were as intended by our design, which was based on the literature. In this study, spectral resolution was manipulated by limiting the number of active electrodes. The experimental conditions (7, 9, 11 and 15 active electrodes) were chosen based on earlier research that had shown a plateau in speech intelligibility for CI users from 7 active electrodes and up (Fishman et al., 1997; Friesen et al., 2001). The effect of spectral resolution on speech intelligibility has been extensively studied (Chatterjee et al., 2010; Fishman et al., 1997; Friesen et al., 2001; Fu et al., 1998; Henry, Turner, & Behrens, 2005; Schwartz et al., 2008; Won, Drennan, & Rubinstein, 2007) and measures of intelligibility are regularly used in both clinical and research settings. The main interest of this study was therefore primarily effects of increased spectral resolution on listening effort when intelligibility is near ceiling, as this is when measures of listening effort can reveal potential benefits that are not directly evident from intelligibility measures.

The results from Experiment 1 showed that for CI users, further increased spectral resolution from 7 active electrodes upwards, did not lead to decreased secondary task RTs. These results are not in line with our expectations based on studies in NH listeners that show improved listening effort for increased spectral resolution. In prior research, we have successfully used this same dual-task paradigm to show improvements in listening effort for increased spectral resolution in NH listeners presented with CI simulated speech, even for conditions that resulted in equal intelligibility (Pals et al., 2013). Other research similarly shows that reduced spectral resolution leads to more effortful speech understanding for NH listeners, as reflected by pupil dilation (Winn et al., 2015). Eye tracking data further suggests that spectral degradation leads to slower speech processing, thus reducing the benefit of sentence context, which can further increase listening effort (Wagner et al., 2016). The speech materials used for the CI users, however, were optimized for use with hearing impaired and CI listeners (van Wieringen et al., 2008): they are spoken with clear articulation and at a slow speaking rate. This may have diminished the detrimental effects of reduced spectral resolution on the effective use of sentence context, as the slower speaking rate may have accommodated for the slower speech processing. In short, the results of this study suggest, that for CI users, specifically in the experimental setting of this study, i.e. when listening to slow-spoken and carefully articulated speech presented without interfering noise and over personal audio cable in a sound isolated booth, improved spectral resolution from 7 active electrodes up does not improve effort as measured by our dual-task paradigm.

Perhaps the lack of effect of spectral resolution on secondary task performance may also be partially explained by a difference in motivation between the NH participants of the previous study and the CI users in the current study. The NH participants were university students participating in a number of studies for course credit and the experimenters observed in these participants a lack of intrinsic motivation to perform well in the experiment. The CI participants on the other hand, were, in our experience, generally very grateful for the improvements scientific research has provided for CI technology, from which they directly benefit, and therefore quite motivated to contribute to scientific research and perform their very best in the experiments. This difference in motivation could have affected the dual-task outcome. The dual-task paradigm reflects effort as a proportion of the total capacity of available resources. However, even if cognitive resources are limited, it has been suggested that the total capacity may be temporarily increased depending on individual strategy,

motivation, and determination to enhance performance by exerting extra effort (Hockey, 1997; Kahneman, 1973; Wingfield et al., 2007). Therefore, if the highly motivated CI users expended extra effort to temporarily increase the cognitive resources available for the dual-task, then, even if the processing load of the primary task increases, it may not necessarily be reflected in performance on the secondary task (Pals et al., 2015, see also Chapter 4).

Another notable difference between the two studies is the ages of the NH versus the CI participant groups. The NH listeners in our previous study (Pals et al., 2013) ranged in age from 19 to 25 years, and the CI listeners in this study ranged in age from 34 to 76 years. Age is known to affect cognitive ability, and could in our CI participant group have resulted in reduced cognitive capacity compared to their younger NH counterparts. However, reduced cognitive capacity alone cannot explain the lack of interaction between the primary listening task and the secondary response time task. Perhaps the problem lies with the nature of the secondary task; a response-time task that requires a fast motor response. Research suggests that cognitive motor control is also affected by age, and more importantly, that divided attention over a cognitive and sensory-motor task, such as in our dual-task paradigm, greatly affects motor control in older adults (Li & Lindenberger, 2002; Seidler et al., 2010). This may explain the fact that the RTs on the secondary task in this study, are around half a second longer than the RTs for the young NH participants in our previous study. This increased cost of divided attention and cognitive motor control may have resulted in a floor effect for the rhyme-judgment task, and could therefore be a reason why the RTs showed no additional effect of improved spectral resolution of the stimuli.

In Experiment 2, a simple, single-task, response-time measure was used as a measure of listening effort; the sentence verification task (Adank et al., 2009; Baer et al., 1993). In addition to the CI users participating in both Experiment 1 and 2, an extra group of young NH participants was recruited for a validation experiment. The sentence verification task is considered a measure of comprehension (accuracy), and speed of comprehension (RTs). A measure of comprehension requires the listener to understand the meaning of the speech and reason about it (Ralston et al., 1991; Wingfield et al., 2007), and may therefore more closely reflect the requirements of everyday verbal communication. In the sentence verification task, the RTs reflect the processing time required to comprehend the speech and judge whether the sentence was true or false. Research suggests that increased listening effort results in longer

processing time required to understand the speech (Gatehouse & Gordon, 1990; Gibbon et al., 1997; Pals et al., 2015; Wagner et al., 2016). We therefore interpret longer RTs as increased listening effort.

The results of Experiment 2 showed improved sentence verification task accuracy scores, i.e. comprehension, with increased spectral resolution for NH listeners at least up to 12 spectral channels, and for CI users up to 11 active electrodes. In our previous study, the dual-task intelligibility results in NH listeners, improved only up to 6 spectral channels (Pals et al., 2013). Other research similarly shows ceiling performance on sentence intelligibility in quiet for NH listeners for spectral resolution of around 5 to 6 spectral channels, sentence intelligibility in noise, however, continues to improve with increased spectral resolution up to 12 to 20+ spectral channels depending on the SNR (Dorman et al., 1998; Dorman, Loizou, & Rainey, 1997; Friesen et al., 2001). Similarly, sentence intelligibility for CI users improves with increased spectral resolution, in quiet up to 4 to 7 active electrodes, and in noise up to 10 active electrodes (Fishman et al., 1997; Friesen et al., 2001). The sentence verification task comprehension scores in quiet appear more similar to the speech intelligibility results in noise reported in these studies. This might suggest that the sentence material used for this sentence verification task is more challenging than the sentences typically used for intelligibility tasks. Alternatively, comprehension, the understanding and reasoning about the heard speech as needed to judge the truth value of a sentence, may be more affected by changes in cognitive processing load than a measure of intelligibility is. As speech comprehension requires further cognitive processing than the first step of speech perception that is measured in an intelligibility task, comprehension is suggested to rely more on cognitive capacity (Just & Carpenter, 1992; Ralston et al., 1991). Comprehension may thus be more affected by changes in listening effort than intelligibility.

In addition to the improvement in comprehension, the NH results showed a clear linear trend of improved RTs up to at least 16 spectral channels. Similar to our previous dual-task study in NH listeners, the sentence verification task RTs show more clearly and convincingly an improvement with increased spectral resolution from 12 to 16 channels, while the accuracy scores, i.e. comprehension, appeared to reach ceiling around 12 channels (see Figure 3). In CI users, on the other hand, both accuracy and RTs appeared to reach ceiling at around the same level of spectral resolution; around 11 active electrodes. The results of Experiment 2

show that, even when the dual-task results for speech in quiet no longer show improved sentence intelligibility or listening effort, the sentence verification task suggests that further improved spectral resolution can still further improve speech comprehension and listening effort, both in NH and CI listeners.

In addition to the potential problems with our dual-task paradigm and the older CI participants as discussed above, the different speech materials used for the dual task and for the sentence verification task may have contributed to the differences in outcomes between the two measures. The speech stimuli used in Experiment 1 were taken from the LIST corpus that is optimized for hearing-impaired and CI listeners (van Wieringen et al., 2008), chosen to allow the CI participants to achieve near ceiling performance on the primary listening task. In Experiment 2, the sentences were spoken by a native Dutch speaking, young-adult male speaker, speaking at normal conversational speed, and therefore likely more challenging to understand for CI users than the speech material used in Experiment 1. Wingfield et al. (2006) suggest that effects on speech comprehension become apparent only after a certain threshold of processing difficulty has been crossed, and therefore the nature of the speech material and task affect the outcome of such tests. Perhaps in Experiment 2, the more challenging speech materials result in a stronger effect of spectral resolution on task performance.

However, the difference in results between the dual task and the sentence verification task may also, in part, be due to the nature of the tasks themselves. In a previous study (Pals et al., 2015), we found a similar difference in effects shown by the dual-task paradigm and a simple verbal RT measure of listening effort, in an experiment with young adult NH participants listening to speech in various noise conditions. In this previous study, both tasks were performed by the same participants, and using the same speech materials. The differences in outcomes between the two tasks can therefore not be attributed to differences in age or motivation of the participants, or to differences in speech materials, suggesting that they must stem from differences between the two measures themselves. In the current study, the difference in outcomes between the dual task and the sentence verification task may also, in part, be due to differences in the nature of the tasks.

Regardless of the reason for the differences between the dual task and the sentence verification task outcomes, the core finding of this study is this: the sentence verification task

has shown improved speech comprehension and listening effort in CI users for improved spectral resolution between 7 – 11 active electrodes, conditions in which speech intelligibility measures typically show no change. The same spectral manipulation in Experiment 1 showed no effect on speech intelligibility and listening effort as measured using the dual-task paradigm, and other research also shows a plateau in speech intelligibility in quiet listening conditions for spectral resolution beyond 7 active electrodes in CI users (e.g. Fishman et al., 1997; Friesen et al., 2001). In other words, the sentence verification task has shown a benefit of spectral resolution, that is not likely to be detected by the clinical speech intelligibility tests and may therefore be a valuable measure to complement the traditional intelligibility measures and reveal some of the cognitive processing underlying speech understanding.

In conclusion, spectral resolution does affect speech comprehension and listening effort in CI users. Even in highly idealized listening conditions, clear speech presented without background noise, through personal audio cable in a sound proof room, the sentence verification task showed both improved speech comprehension and listening effort for increasing spectral resolution up to 11 active electrodes. This finding shows both the benefit of increased spectral resolution for CI users even when this benefit is no longer evident from speech intelligibility measures, and thus also the added value of a measure such as the sentence verification task to complement traditional measures of intelligibility to uncover such potential benefits. Our specific dual-task paradigm may not be the method of choice for measuring listening effort in CI users. The sentence verification task shows clear effects of changes in spectral resolution on both speech comprehension and listening effort, the task is easier to explain to participants, easier to perform, and easier to implement than the dual task, making it an attractive method for use in both research and for clinical purposes.

