Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test

Verena N. Uslar[a),b)]
Department of Medical Physics and Acoustics, Carl von Ossietzky University, Oldenburg, Germany

Rebecca Carroll,[b)] Mirko Hanke, Cornelia Hamann, and Esther Ruigendijk[b)]
School of Linguistics and Cultural Studies, Carl von Ossietzky University, Oldenburg, Germany

Thomas Brand[b)] and Birger Kollmeier[b)]
Department of Medical Physics and Acoustics, Carl von Ossietzky University, Oldenburg, Germany

(Received 6 November 2012; revised 25 July 2013; accepted 2 August 2013)

To allow for a systematic variation of linguistic complexity of sentences while acoustically controlling for intelligibility of sentence fragments, a German corpus, Oldenburg linguistically and audiologically controlled sentences (OLACS), was designed, implemented, and evaluated. Sentences were controlled for plausibility with a questionnaire survey. Verification of the speech material was performed in three listening conditions (quiet, stationary, and fluctuating noise) by collecting speech reception thresholds (SRTs) and response latencies as well as individual cognitive measures for 20 young listeners with normal hearing. Consistent differences in response latencies across sentence types verified the effect of linguistic complexity on processing speed. The addition of noise decreased response latencies, giving evidence for different response strategies for measurements in noise. Linguistic complexity had a significant effect on SRT. In fluctuating noise, this effect was more pronounced, indicating that fluctuating noise correlates with stronger cognitive contributions. SRTs in quiet correlated with hearing thresholds, whereas cognitive measures explained up to 40% of the variance in SRTs in noise. In conclusion, OLACS appears to be a suitable tool for assessing the interaction between aspects of speech understanding (including cognitive processing) and speech intelligibility in German.
 [http://dx.doi.org/10.1121/1.4818760]

PACS number(s): 43.71.Sy, 43.71.Gv, 43.71.Rt, 43.72.Dv [MAS] Pages: 3039–3056

I. INTRODUCTION

Understanding speech is a complex human ability that relies both on bottom-up, sensory processing, mainly limited by the amount of acoustical information contained in a (noisy) speech signal, and on top-down, cognitive processing, which is limited by processing capacity and working memory constraints, among others. Psycholinguists typically concentrate on the cognitive aspects and measure differences in speech processing that are related to cognitive capacity limitations and/or linguistic complexity. In contrast, audiologists typically concentrate on sensory processing as they are interested in an individual's hearing disorders. Speech reception thresholds (SRTs) measure the amount of sensory information (expressed as the speech level in quiet or the signal-to-noise ratio in background noise) required for an individual to reach a certain performance level, e.g., 50 or 80% correctly repeated speech items. Ignoring cognitive effects in speech audiometric tests may, however, lead to wrong conclusions about the speech reception capacities in the individual patient. Conversely, ignoring the audiologically defined acoustical information limitations in speech materials may lead to incorrect conclusions about the relative contribution of linguistic structure and cognitive processes. The aim of the current study therefore is to provide and verify a tool for quantifying the interaction between top-down cognitive and bottom-up sensory processing in speech perception. The Oldenburg linguistically and audiologically controlled sentences (OLACS) (http://www.aulin.uni-oldenburg.de, Last viewed 08/15/2013) systematically vary linguistic complexity while controlling for several relevant types of acoustic information. Homogeneous acoustical information transfer across and within each sentence type was ensured by a careful selection and evaluation process, which was designed to accommodate the unique structure of the sentence material. At the end of the selection process the material was reduced from initially over 1000 sentences to 280, which now comprise the OLACS corpus for speech intelligibility measurements.

A. Linguistic complexity in psycholinguistic research

Experimental studies on language processing have manipulated different parameters to vary linguistic complexity (see, for instance, Gordon *et al.*, 2001; Fanselow *et al.*, 1999). These parameters are mostly derived from assumptions based on linguistic theory.[1] In those studies, an increase in the number of comprehension errors, an increase in reaction times, or a change in electrophysical correlates for complex structures compared to structurally simpler sentences is often interpreted as the result of a greater processing "load." Of the different structural factors that have been

[a)]Author to whom correspondence should be addressed. Electronic mail: verena.uslar@uni-oldenburg.de
[b)]Also at: Cluster of Excellence "Hearing4all".

identified in the literature to affect the ease of processing, we focus on the following three.

(1) **Word order.** German allows for a relatively free word order. The position of a word bearing a particular grammatical function (e.g., subject, object, adverbial) may vary without changing the meaning of the sentence. Like many other languages, German nevertheless shows a general preference for subject-before-object word order over configurations in which the object precedes the subject (e.g., Bader and Meng, 1999; Gorrell, 2000). For example, (1a) below is considered a preferred or *canonical* structure in comparison to (1b). Note that (1b), unlike a passive such as (1c) does not truly change the meaning of the sentence. The hugger (= subject) is in the nominative case in (1a) and (1b), but not in (1c). Note also that the non-canonical order in (1b) is highly restricted to specific discourse contexts (Fanselow *et al.*, 2008).

(1a) *Der Vater umarmt seinen Sohn.*
The$_{\text{NOMINATIVE}}$ father hugs his$_{\text{ACCUSATIVE}}$ son.
The father hugs his son.

(1b) *Den Vater umarmt der Sohn.*
The$_{\text{ACCUSATIVE}}$ father hugs the$_{\text{NOMINATIVE}}$ son.
The son hugs the father.

(1c) *Der Vater wird von dem Sohn umarmt.*
The$_{\text{NOMINATIVE}}$ father is by the$_{\text{DATIVE}}$ son hugged.
The father is hugged by the son.

(2) **Embedding.** The existence of embedded relative and subordinate clauses yields more structure on a syntactic, as well as on a semantic level, which arguably leads to higher processing costs. For example, "the boy laughs" yields less structure and less semantic information than "the boy who is seeing the woman laughs." With certain relative clause constructions in particular, the (negative) processing effects of embedding can be augmented further by manipulating word order within the embedded clause (e.g., Bader and Meng, 1999; Gordon *et al.*, 2001). The most prominent example found in the literature is the distinction between subject (2a) and object (2b) relative clauses (see examples below). Whereas in the canonical word order the subject precedes the object as in (2a), in the non-canonical order the object precedes the subject in (2b).

(2a) The boy who really liked the girl blushed.

(2b) The girl whom the boy really liked blushed.

(3) **Ambiguity.** Formal identity of functionally distinct words (e.g., lexical ambiguity of verbs and nouns as in "the old man the boat" or ambiguously case-marked articles) can lead to temporary uncertainties during processing, for instance with regard to the grammatical function and/or semantic role of constituents in a sentence (see overview by Altmann, 1998). In German, so-called reanalysis effects can be found in word order ambiguities ("who did what to whom?"). For example, the definite article most visibly marks the grammatical case of a noun phrase. However, the inflectional paradigm of German articles contains a number of ambiguous word forms (such as *die*, "the," which can be either nominative or accusative for singular female nouns or gender neutral plural in all cases). In sentences like (3a), the ambiguity of the article *die*, "the," can lead to temporal uncertainty as to whether a constituent (the pretty woman) is marked nominative (usually the subject of a sentence) or accusative (usually the direct object; Bader and Meng, 1999).

(3a) *Die hübsche Frau mag der Mann besonders gern.*
The$_{\text{AMBIGUOUS}}$ pretty woman likes the$_{\text{NOMINATIVE}}$ man particularly well.
It is the pretty woman whom the man likes particularly well.

In temporally ambiguous sentences such as (3a), the parser has to reanalyze the initial subject interpretation of *die hübsche Frau* at a later point, namely, the point of disambiguation, here: *der Mann*, which is unambiguously marked as nominative and therefore must be interpreted as the subject of the sentence. This kind of reanalysis has been argued to come at increased processing cost and higher error rates (e.g., Frazier and Rayner, 1982).

Psycholinguistic studies have provided strong evidence that syntactically complex sentences are more difficult to process than syntactically simpler sentences especially under adverse listening conditions; this can be observed, for instance, as an increase in reaction time or error rate (e.g., Wingfield *et al.*, 2006; Carroll and Ruigendijk, 2013). For example, Tun *et al.* (2010) showed that when participants had to listen to sentences and then respond to a true/false comprehension question, response latencies were significantly higher for complex sentences than for simpler sentences. In addition, speech comprehension is hindered by semantic manipulations, for instance, by using homophones [words with multiple meanings; Rodd *et al.* (2005)]. Those studies showed that sentence complexity indeed seems to prolong sentence processing. Thus, since slowing down of processing seems a good indication of processing difficulties in complex sentences, we used response latency measurements as an additional, easily obtainable measure for the verification of the final OLACS material.

B. Linguistic complexity and audiology

Modern speech audiometry measures either the rate of correctly perceived words or the SRT, using audiologically controlled items that are very similar in their item-specific reception thresholds (reviewed by Vlaming *et al.*, 2011). This involves a combination of an auditory bottom-up processing task and a cognitive (working memory) task, which does not necessarily involve understanding, but rather simply recognition of the words that were heard in sentences. Understanding, based on a structural analysis, however, may play a role as well. Although not all words are necessarily understood correctly, integrating the words that are heard into a meaningful sentence seems to facilitate word recall (Uslar *et al.*, 2011).

Hence, we wished to further clarify to what degree linguistic complexity influences speech reception, and whether a closer control of linguistic complexity is required for speech audiometric tests. The material used in existing speech audiometric intelligibility tests, consisting of short, meaningful sentences, like the German-language Göttingen

sentence test (GÖSA) (Kollmeier and Wesselkamp, 1997) and its equivalent in different languages (see, for instance, Plomp and Mimpen, 1979, Ozimek *et al.*, 2009, and others) has not been controlled for linguistic complexity (sentence structure, word frequency, ambiguity, etc.). For the GÖSA, Uslar *et al.* (2011) found a small effect of linguistic complexity on speech recognition for young listeners with normal hearing, whereas speech recognition in older listeners with normal or impaired hearing was not affected by linguistic complexity. The findings were attributed to possible differences in strategies used by older listeners and their more extensive experience with understanding speech in adverse listening conditions. Uslar *et al.* (2011) thus concluded that speech audiometric tests that control and manipulate linguistic complexity may have the potential to reveal interactions between central language processing capabilities and speech recognition scores in a clinical population.

Pichora-Fuller (2008) showed that older adults benefit at least as much as younger adults from a supportive context. They rely more on redundant cues in the speech signal or internally stored knowledge about common sentences or sentence structures, thus overcoming potential processing difficulties by employing top-down processing. Thus, semantically unpredictable sentences with varying syntactic complexity may produce different results from those reported by Uslar *et al.* (2011), because the strategies used by experienced listeners may not work for these sentences.

One alternative for measurements with unpredictable sentences are speech audiometric intelligibility tests using matrix sentences (e.g., Wagener *et al.*, 1999, and its equivalent in different languages; Vlaming *et al.*, 2011). Another type of test material are SUS-test like sentences (Benoît *et al.*, 1996; mainly used for the evaluation of the intelligibility of text-to-speech systems at sentence level) or material like the one introduced by Bolia *et al.* (2000). All these materials contain sentences with correct and fixed syntax, but little semantic predictability within each sentence because of little or no supportive context. For example, the OLSA (Oldenburger Satztest) sentence, *Peter malt zehn nasse Sessel* ("Peter draws ten wet armchairs") (Wagener *et al.*, 1999) is considered as a semantically very unusual sentence with little supportive context. Therefore the correct recognition of *Sessel,* "armchair," would probably not facilitate the recognition of the words *malt*, "draws," or *nasse*, "wet," and vice versa. The OLSA sentences differ only minimally in their linguistic complexity (i.e., verb argument structure and word frequency).

C. Evaluation of speech material for audiometric tests

Evaluation of new material for speech intelligibility tests follows a certain procedure, depending on the starting point for the material and the specific requirements for the respective material (see, for instance, Wagener *et al.*, 1999; Plomp and Mimpen, 1979). For example, evaluation procedure may begin with a stock of 100 sentences, which are then carefully evaluated regarding the sentence specific intelligibility function [with measurements at different signal to noise ratios (SNRs)] and afterwards optimized by slight corrections to the sound level of specific words or whole sentences.

However, there are different reasons to deviate from this much-practiced method. If, for instance, the initial corpus is very large, as in our case, a somewhat reduced effort in the assessment of the sentence specific intelligibility functions followed by a systematic rejection of outliers can be an appropriate and feasible method to assure high homogeneity in acoustic information content across test items. After all, the ultimate goal of homogenization is to achieve high comparability of test items in terms of SRT outcome.

For OLACS we introduced an additional control parameter, namely, linguistic complexity. When this is done, it has to be taken into account that an adjustment of the speech sound pressure levels might compensate for the effect of linguistic complexity, thus hampering the separation of linguistic and auditory effects. For that reason we did not adjust speech levels but selected those sentences that were close to the specific mean of the measured SRT_{80} (i.e., the speech reception threshold at which 80% of the presented stimuli are repeated correctly) for the specific sentence type. This method of rejecting outliers was successfully implemented in the optimization of intelligibility tests using everyday sentences by, for instance, Ozimek *et al.* (2009) or Kollmeier and Wesselkamp (1997).

In summary, more appropriate test materials are needed that provide parametric control of linguistic complexity while still exhibiting the high homogeneity in intelligibility across test items typical in modern speech audiometric tests. The design of the OLACS, described below, could serve this purpose: Its design objectives are defined by both linguistically and audiologically important parameters. The sentences include several types of linguistic complexity with a fixed sentence structure and low predictability for each respective type.

D. Cognitive effects related to speech processing

Several studies have shown that, apart from sensory-acoustical factors measured in audiology, cognitive abilities seem to be relevant for speech processing as well. In particular, working memory capacity has been attributed a central role (e.g., Just and Carpenter, 1992; Caplan and Waters, 1999; Zekveld *et al.*, 2011). In audiological research, measures of working memory capacity and attention have been found to correlate with speech intelligibility results under certain conditions (review by Akeroyd, 2008, and see, e.g., Rudner *et al.*, 2011). Furthermore, cognitive measures seem to be useful in predicting the outcome of hearing aid fittings, since they provide prior information about the potential benefit and acceptance of hearing aids (e.g., Lunner *et al.*, 2009). This study therefore includes cognitive measures to further validate the test with an independent measure of individual cognitive processing capacities. We compared the SRT_{80} values with experimental measures estimating the individual working memory storage capacity (word span forward; see Schuchart, 2008), the ability to manipulate the content of one's working memory (digit span backwards; the revised German version of the Wechsler Adult Intelligence scale; Tewes, 1991), and a measure for attention and susceptibility to interference (Stroop test; Stroop, 1935).

We chose two types of working memory tests, because remembering the content words of a sentence and then manipulating spoken input are both important for speech processing and understanding. Additionally, we employed a variation of the Stroop test as a measure of general attention and a measure of participants' susceptibility to interference (i.e., the ability of the participant to ignore additional confounding visual information unrelated to the actual visual task, which has been argued to be independent from span measures; May *et al.*, 1999). Susceptibility to interference is widely believed to be a supra-modal cognitive resource (e.g., Reisberg, 2007). Thus, measures in the visual mode are expected to translate well to auditory tasks. The susceptibility to interference was expected to play a role in noisy listening conditions, since noise can also generally be viewed as a kind of interference, and general attention to the current task should facilitate performance, in our case understanding speech in difficult listening conditions (see, for instance, Rönnberg *et al.*, 2010).

II. METHODS

The first part of this section describes the design and implementation of the test corpus, while the rest describes the verification of its properties using SRT_{80} and response latency measurements, as well as cognitive tests. The validation of the test materials developed here, with respect to their perceptual properties (i.e., SRT_{80} and response latency for different sentence complexities), was then performed in quiet, in stationary, and in fluctuating noise. The noise conditions were selected as best representing standard speech audiometrical configurations: Stationary noise represents a well-defined deterioration of the sensory information contained in the speech signal, whereas fluctuating noise provides an extra challenge to the listeners' cognitive functions when combining the information from unmasked speech portions ("glimpses" during short noise gaps, see, for instance, Rhebergen *et al.*, 2006). Although this possible dip-listening in fluctuating noise leads to lower SRTs when compared to stationary noise, to reach these lowered thresholds, one has probably employed more cognitive resources. Thus, getting better thresholds in fluctuating noise comes with a higher effort, i.e., a need for a stronger cognitive contribution on the listener than in stationary noise or in quiet (e.g., Rudner *et al.*, 2011) and may therefore explain parts of the larger inter-individual variability in SRT_{50} for fluctuating noise compared to stationary noise or quiet (Bronkhorst, 2000; Wagener *et al.*, 2006).

A. Design and implementation of the German OLACS

In the following the development of the OLACS corpus is described: first, the construction principles applied, and second, the process of narrowing down the number of sentences from nearly 1000 potential candidates to the 280 sentences that now make up the OLACS corpus.

1. Construction principles

For the OLACS material we combined the three factors of linguistic complexity, that is, word order, embedding, and ambiguity since these phenomena have been examined frequently with relatively consistent results as described above.

TABLE I. Examples for the seven types of sentences employed in the OLACS in the original German version and word for word English translations. Small capitals indicate the earliest possible point of disambiguation in each sentence type. Nom (nominative), acc (accusative), and amb (ambiguous) indicate the case of the words. sg indicates singular forms, pl indicates plural forms. Verbs are either in their third person singular (3sg) or third person plural (3pl) form. Vertical lines indicate the points where the sentences were cut to create the fragments used for evaluation phase I (see Sec. II A 3).

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7
1. SVO	DER	kleine	Junge	grüsst	den	lieben	Vater.
	THE	little	boy	greets	the	nice	father
	nom	*nom*		*3sg*	*acc*	*acc*	
2. OVS	DEN	lieben	Vater	grüsst	der	kleine	Junge.
	THE	nice	father	greets	the	little	boy
	acc	*acc*		*3sg*	*nom*	*nom*	
3. ambOVS	Die	liebe	Königin	grüsst	DER	kleine	Junge.
	the	nice	queen	greets	THE	little	boy
	amb	*amb*		*3sg*	*nom*	*nom*	
4. SR	Der	Bauer,	DER	die	Lehrer	fängt,	lacht.
	the	farmer	WHO	the	teachers	catches	smiles
		sg	*nom*	*amb*	*pl*	*3sg*	*3sg*
5. OR	Der	Bauer,	DEN	die	Lehrer	fangen,	lacht.
	the	farmer	WHOM	the	teachers	catch	smiles
		sg	*acc*	*amb*	*pl*	*3pl*	*3sg*
6. ambSR	Die	Bauern,	die	die	Lehrerin	FANGEN,	lachen.
	the	farmers	who(m?)	the	teacher	CATCH	smile
		pl	*amb*	*amb*	*sg*	*3pl*	*3pl*
7. ambOR	Die	Bauern,	die	die	Lehrerin	FaNGT,	lachen.
	the	farmers	who(m?)	the	teacher	CATCHES	smile
		pl	*amb*	*amb*	*sg*	*3sg*	*3pl*

Thus, we constructed seven sentence types which fall into one of two main categories (see Table I for examples; for sound examples please refer to http://www.aulin.uni-oldenburg.de/49349.html, Last viewed 08/15/2013):

Category one: sentences with verb-second structure.

(1) Sentences containing a transitive verb (i.e., a verb which requires a subject and one object, for example "hug"), with canonical subject-verb-object word order and unambiguous allocation of grammatical functions and semantic roles (SVO in Table I).
(2) Sentences containing a transitive verb, with object-first word order and unambiguous allocation of grammatical functions and semantic roles (OVS). See example (1) for a distinction between canonical SVO and non-canonical OVS structures.
(3) Sentences containing a transitive verb, with object-first word order and ambiguous case marking on the object noun phrase (ambOVS). Refer to example (3) above for further explanation.

Category two: sentences with relative-clause structure [see sentence example (2) for an example in English].

(4) Sentences containing an intransitive verb (i.e., a verb that only requires a subject, but no object, for example "laugh") and a centrally embedded subject relative clause (relative pronoun is the subject of the embedded clause) with unambiguous allocation of grammatical functions and semantic roles within the relative clause (SR).

(5) Sentences containing an intransitive verb and an embedded object relative clause (relative pronoun is the object of the embedded clause) with unambiguous allocation of grammatical functions and semantic roles within the relative clause (OR).
(6) Sentences containing an intransitive verb and an embedded subject relative clause with initial function/role ambiguity (ambSR).
(7) Sentences containing an intransitive verb and an embedded object relative clause with initial function/role ambiguity (ambOR).

We expected non-canonical word order to negatively influence the different outcome measures used throughout the Deutsche Forschungsgemeinschaft funded AULIN project (in whose scope the OLACS material was developed), i.e., higher SRTs and longer response latencies. Introducing additional ambiguity was expected to result in even worse outcomes.

All sentences in the corpus contain seven words. These words were controlled for a number of linguistic variables, namely, the number of syllables, lexical frequency of occurrence for the verbs, and the possibility of being depicted. Verb-second transitive sentences were constructed to contain between 11 and 13 syllables and the relative clause sentences were designed to have 10 or 11 syllables.

The 27 transitive verbs used in the embedded relative clauses and in the verb-second sentences do not comprise any particle verbs (e.g., *ansehen*, "to look at," in which the particle *an* is split off the verb stem in a verb-second clause: *er sieht sie an*, "he looks at her"). Each verb contains 1–4 syllables and belongs to a frequency class between 10 and 18, according to the Leipziger Wortschatz corpus (Biemann *et al.*, 2004). Items of class n occur 2^{-n} times as often as the most frequent word in the corpus, i.e., the definite article *der*, "the" (nom;masc). All verbs employed have only one argument structure,[2] since the number of argument structures is known to influence sentence comprehension (Shapiro *et al.*, 1987). For the intransitive main clauses that contain the embedded relative clauses, verbs were chosen that describe actions or events that can possibly take place simultaneously with other actions; most of these verbs denote bodily reactions like laughing, sweating, smiling, or crying.

All nouns are between 1 and 3 syllables long and the collection spans a range of frequency classes from 8 to 17. The set of nouns comprises only words denoting animate, and more specifically, only human or anthropomorphous entities. In order to achieve an equal number of words across sentence types, a matching adjective was added to each of the two nouns in the simple transitive sentences. These of course do contribute to the amount of information contained in the transitive sentences. Last, for each sentence, we considered whether the entities and the action or event denoted by the sentence could be depicted in an easily recognizable fashion.

2. Plausibility test

To ensure that the intelligibility of a given sentence was not influenced by the underlying semantic plausibility of the sentence we ran a large-scale plausibility study. The goal was to ensure that the two arguments (agent and patient or subject and object) are, in principle, interchangeable; that is, they should be equally likely to be interpreted as the subject (or object) argument in order to avoid any semantically driven interpretation preferences that could lead to processing advantages for either a canonical or non-canonical reading (e.g., Mak *et al.*, 2002; Bornkessel *et al.*, 2005). For instance, the argument *Polizist*, "policeman," in sentence (4a) below is semantically more likely to be the agent and subject than *Dieb*, "thief," since policemen are usually catching thieves and not vice versa. The noun *Kapitän*, "captain," in sentence (4b) is semantically just as likely to be the subject as *Matrose*, "sailor."

(4a) *Der Polizist fängt den Dieb.*
The policeman catches the thief.
(4b) *Der Kapitän weckt den Matrosen.*
The captain wakes the sailor.

Reasons for this difference in subject likelihood between (4a) and (4b) are the semantics of the verb and the two respective nouns.

Plausibility was tested in the form of questionnaires in which we presented all 896 sentences (128 items per sentence type) in the original order, and their counterparts, in which the arguments were interchanged, as exemplified in (5) below.

(5a) *Der kleine Junge umarmt den lieben Vater.*
The little boy hugs the dear father.
(5b) *Der liebe Vater umarmt den kleinen Jungen.*
The dear father hugs the little boy.

This amounted to a total of 1792 critical items, which were randomly distributed across 43 questionnaire lists. Each questionnaire consisted of 63 sentences (42 critical and 21 nonsense filler items) and was filled out by at least 10 participants, which amounted to about 500 undergraduate students at the Carl von Ossietzky (CvO) University, Oldenburg, all native speakers of German. We ensured that the same verb only occurred once or twice within one questionnaire. The participants' task was to rate the semantic plausibility (but not grammatical correctness) for each sentence on a six-point scale that was based on the German school grading system (1 = very plausible, 6 = not at all plausible). The study, and all experiments described below, was approved by the ethics commission of the CvO University, Oldenburg.

We compared the difference in the scores of each sentence pair within one condition. Pairs whose score difference exceeded 2 standard deviations (s.d. = 0.51) from the mean of difference (0.65) between sentence pairs were excluded from the corpus (that is, all pairs whose score difference exceeded 1.68). Outlier sentences per condition were eliminated. Six verbs deviated in their results across sentences and conditions, such that their averages exceeded a score of 3.5 and these verbs were therefore also excluded from the corpus. After thus considerably reducing variation between sentences in each condition, our corpus shrank to a total of 680 sentences.

3. Homogenization of the intelligibility of the sentence material

The remaining 680 sentences were recorded in a large sound-proof chamber. The speaker was a female speech therapist with classical vocal training. The intent was to generate natural-sounding sentences, spoken at a moderate rate, with no strong emphasis on any words or parts of words. We chose the best out of five or six recordings for each sentence. If none of the recordings met these requirements regarding the sound and/or timing, we eliminated the sentence from the test corpus. Thus our corpus was reduced to 568 sentences.

A primary design goal for modern speech audiometric tests is a high homogeneity in test-item-specific intelligibility, in order to maintain a high slope of the discrimination function and a high efficiency of the test in measuring speech reception thresholds (see Kollmeier and Wesselkamp, 1997; Wagener *et al.*, 1999; Vlaming *et al.*, 2011). To equalize the sentence-specific intelligibility across all test sentences (and relevant segments for each of the seven sentence types, see below), three successive intelligibility evaluation measurements were done with 42 listeners with normal hearing. They were all students at the CvO University of Oldenburg and received 10 Euros per hour for their participation.

All experiments took place in a sound-insulated booth ($1.87 \times 1.53 \times 2.51$ m; $D \times W \times H$). A standard PC equipped with an RME Digi96/8 PAD sound card was used to generate stimuli at a sampling frequency of 44 100 Hz. Signals were then processed by an RME ADI-8 Pro eight-channel D/A converter, amplified by a Tucker-Davis HB7 headphone driver. Sentences or sentence fragments were presented via HDA200 Sennheiser headphones that were free-field equalized using a finite impulse response filter with 801 coefficients according to DIN norm (DIN EN 389-8, 2004). All technical equipment was situated outside the experimental booth, except for the headphones and the touch screen computer monitor used by the experimenter. Listeners heard the sentences and sentence fragments in stationary noise [65 dB sound pressure level (SPL)] with the long-term frequency spectrum of the speaker's voice, created by overlapping 30 tracks, each consisting of the entire randomly overlapping speech material. They then repeated the words they heard after the end of each sentence. Correctly repeated words were scored by the experimenter sitting in the soundproof chamber with the participant. This general setup and procedure was the same for all audiological tests used in this study.

a. Evaluation phase I. In general, presenting words in a sentence context is regarded to be useful for the listener (see, e.g., Miller *et al.*, 1951, Pickett and Pollack, 1963; Boothroyd and Nittrouer, 1988; Plomp, 2002). In the case of the OLACS material, we hypothesized that—at least in the case of ambiguous and noncanonical sentences—the sentence context might be detrimental. Herein lies one of the major problems that arose during the homogenization of the material. As mentioned above, a main goal in creating new sentence material suited for audiological purposes is keeping the acoustical information transfer content as homogenous as possible throughout the entire material. In the case of OLACS, we expected linguistic complexity to be detrimental to speech intelligibility scores. We thus had to ensure that all differences between sentence types were due to linguistic complexity and not to differences in the acoustic representation.

Therefore, in evaluation phase I, only sentence fragments were presented, in order to measure the intelligibility independently of the linguistic complexity of the complete sentences, thus reducing both positive and detrimental context effects and focusing on acoustical cues only. Each sentence fragment was presented at a fixed SNR of −7 dB SNR (i.e., the unweighted SNR at which, according to preliminary measurements, approximately 50% of the words were understood correctly). Intelligibility was measured using word scoring for fragments of 568 sentences, which resulted from cutting each sentence into three pieces. Verb-second sentences were cut directly before and after the verb (see vertical lines in Table I), yielding fragments such as *Der nette Elefant*, "the nice elephant," *umarmt*, "hugs," and *den kleinen Igel*, "the small hedgehog." Relative clause sentences were cut at both commas (see vertical lines in Table I), yielding fragments like *Der Zauberer*, "the wizard," *der die Zwerge interview*, "who interviews the dwarfs," and *lacht*, "laughs." All sentences in which one or more words were never understood when presented in a fragment were marked as possible candidates for rejection.[3] Thus, at this point the material was not reduced and all 568 sentences were then used in evaluation phase II.

b. Evaluation phase II. In evaluation phase II the intelligibility of the remaining 568 sentences was measured presenting the complete sentences, again at an SNR of −7 dB SPL. Based on the results of evaluation phase II, the mean intelligibility and the standard deviation for each sentence type were calculated. Sentences that deviated in intelligibility by more than two standard deviations from the mean of the respective sentence type average were discarded. Rejection candidates of evaluation phase I were subsequently taken into account as well. This procedure resulted in a total of 360 sentences with optimized homogeneity within each sentence type. In addition, comparing intelligibility of complete sentences (phase II) with that of sentence fragments (phase I), allowed us to determine the effect of syntactic differences between the different sentence types.

An effect of linguistic complexity on speech intelligibility could already be observed here by comparing the recognition rates of the fragments from phase I to the recognition rates of the same fragments when presented within a sentence from phase II. Note that during evaluation phases I and II, the fragments in isolation and in sentence context, respectively, did not have exactly the same SNR. For better comparison in evaluation phase I, all fragments were presented at an SNR of −7 dB (based on fragment RMS values), whereas in evaluation phase II the complete sentences were presented at a fixed SNR of −7 dB (sentence RMS values), so that each fragment was presented at a slightly different SNR. This mismatch was corrected using a simple linear extrapolation. For each fragment the intelligibility ($\text{intel}_{\text{frag}}$)

of evaluation phase I was extrapolated to the specific SNR ($SNR_{inSentence}$) used for this fragment in evaluation phase II. This extrapolated $intell_{corr}$ was calculated as

$$intell_{corr}[\%] = intell_{frag}[\%] + 10\left[\frac{\%}{dB}\right] \times (-7[dB] - SNR_{inSentence}[dB]). \quad (1)$$

By first determining the difference between the −7 dB SNR at which the fragment was presented during evaluation phase I and $SNR_{inSentence}$ and then assuming a mean slope for the psychometric function of 10% per dB (this slope is an estimate for our speech material when only fragments are presented[4] and is normally found for short sentences and words in general; see, for instance, MacPherson and Akeroyd, 2012), the theoretically expected intelligibility value can be estimated by multiplying this difference with 10%/dB and adding that to $intell_{frag}$.

Figures 1 and 2 display the mean intelligibility and standard deviations across words for evaluation phase I (dashed gray line and filled gray circles) and for evaluation phase II (black line and filled black squares). Furthermore, Figs. 1 and 2 display the corrected intelligibility for fragments specifying the estimated intelligibility for the fragments according to Eq. (1) when they had been presented at the same level as they were presented at in a whole sentence with a global SNR of −7 dB. The upper panel contains the results for the verb-second sentences and the lower panel contains the results for relative clause sentences.

Figures 1 and 2 contain four important messages.

(1) Comparison between fragments of the same type (gray filled circles; e.g., 1a, 2c, and 3c in the upper panel, or 1c–4c in the lower panel) reveals no noticeable differences in intelligibility. Thus, if level differences are partialed out, intelligibility is constant within one fragment type, indicating that the speech material is homogenous across sentence types.

(2) Comparison of intelligibility of fragments presented at −7 dB SNR (filled gray circles) and estimated fragments intelligibility (extrapolated to the level of evaluation phase II, open gray circles) reveals that the correction was mostly relevant at the end of the sentences. This is likely due to the fact that the speech level naturally decreases towards the end of an utterance. In most cases, the last fragments of each sentence were presented at a higher SNR during evaluation phase I than during evaluation phase II. Evaluation phase I thus overestimates the intelligibility of the last fragment, and in the case of relative clause sentences, the second fragment as well.

(3) Comparison of the intelligibility between sentences types with full sentence presentation (black lines and squares) reveals the expected effect of sentence structure on intelligibility (compare, for instance, the characteristics of the black curve for SVO and OVS sentences in the upper panel or fragments 1a and 2a with 3a and 4a in the lower panel). Intelligibility generally drops for fragments embedded in a more complex sentence structure (OVS versus SVO) or for fragments containing more complex word forms (i.e., plural form in fragments 3a and 4a; in line with findings by Carroll and Ruigendijk, 2013).

(4) Comparison of the open gray circles and black squares reveals the estimated context effect of the OLACS material. In most cases, presenting the fragments in a sentence context seems to be helpful, as the intelligibility for whole sentences lies above the estimated intelligibility for the level-corrected fragments. This is in line with the literature (e.g., Miller *et al.*, 1951, Boothroyd and Nittrouer, 1988). However, in ambOVS sentences, context seems to have a detrimental effect. The estimated intelligibility for the corrected fragments 3c lies above the intelligibility of the same fragments when presented in a sentence context.

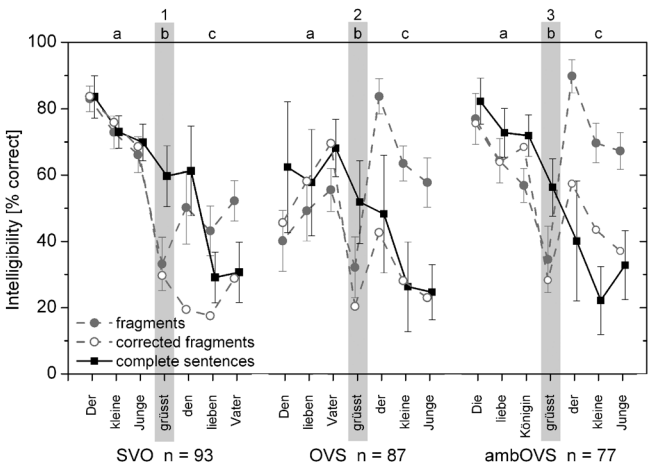

FIG. 1. Mean recognition rates and standard deviations in percent, averaged across listeners for each word position for all verb-second sentences when either fragments (evaluation phase I, with gray lines and circles) or complete sentences (evaluation phase II, with black lines and squares) were presented at an SNR of −7 dB. Open gray circles indicate the corrected intelligibility after extrapolation of the SNR difference between sentence and fragment presentation. The three fragments sentences were cut into are denoted by the letters a–c, the shaded gray area shows the extent of the middle fragment.

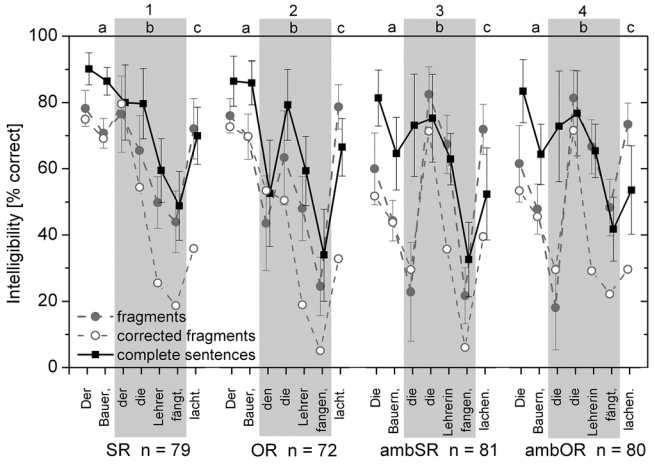

FIG. 2. Mean recognition and standard deviations in percent averaged across listeners for each word position for relative clause sentences when either fragments (evaluation phase I, with gray lines and circles) or completed sentences (evaluation phase II, with black lines and squares) were presented at an SNR of −7 dB. Symbols and shading as for Fig. 1.

In summary, the number of sentences was reduced from 568 to 361 in this evaluation phase. In addition, these initial results suggest a measurable effect of linguistic complexity on speech intelligibility, as evidenced, for instance, by the poorer overall intelligibility of OVS sentences compared to SVO sentences. Additionally, the results provide a clear difference between fragmental intelligibility and intelligibility of a complete sentence presented acoustically. This is confirmed by the fact that the variance across listeners was smaller when complex fragments were presented than when these fragments were presented in the context of a complex sentence. This indicates that listeners may only receive an additional, individually variable benefit in this kind of test if they are able to cope with the complex sentence structure and possibly truly understand the content.

c. Evaluation phase III. To calculate sentence-specific discrimination functions and to exclude those sentences that deviated mostly from the average of the respective sentence type, evaluation phase III determined the recognition rates at two sentence-specific SNRs for the remaining 361 sentences. With the three intelligibility scores per sentence collected for three different SNRs in evaluation phase II and III a sentence specific discrimination function was calculated. The discrimination function was fitted by maximizing the likelihood

$$l(p(L, L_{50}, s_{50})) = \prod_{k=1}^{m} p(L_k, L_{50}, s_{50})^{c(k)} \times [1 - p(L_k, L_{50}, s_{50})]^{1-c(k)}, \quad (2)$$

with m being the total number of words presented to the listener, with $p(L,L50, s50)$ being the discrimination function, and with $c(k) = 1$ if the word k was repeated correctly and $c(k) = 0$ if the word k was not repeated correctly. The parameters L_{50} and s_{50} were varied until the logarithm of the likelihood [i.e., $\log(l(p(L, L_{50}, s_{50})))$] was maximal (see also Brand and Kollmeier, 2002). For the remaining sentences of each sentence type the mean SRT and slope were calculated. In a first step all sentences whose SRT differed more than two standard deviations from the mean SRT of the respective sentence type were excluded. In a second step all sentences with an intelligibility function slope deviating more than one standard deviation from the mean slope of the respective sentence type were rejected. Last, the remaining sentences were ranked by the amount by which the SRTs deviated from the mean SRT of the respective sentence type. Sentences with the largest deviations were excluded until 40 sentences remained in each sentence type. Thus, the seven sentence types of the OLACS corpus comprise a total of 280 sentences.

The resulting mean SRTs stayed constant in each sentence type (mean SRT over all sentence types −8.3 dB SNR; with a maximum of −7.2 dB SNR for OVS sentences and a maximum of −9.5 dB SNR for SR sentences). Standard deviations of SRTs were reduced by about 0.4–1.2 dB in each sentence type. The mean slope was slightly raised from 17.3 to about 17.9 %/dB. Standard deviations for the slope decreased by about 2%/dB to 5.5%/dB.

Figure 3 displays the homogenization outcome based on the speech intelligibility measurements in the three evaluation phases (upper panel verb-second sentences; lower panel relative clause sentences). The standard deviations for the correct repetitions are depicted for each word position measured in evaluation phase II (complete sentences, open bars) after averaging over sentences. The variance in each sentence type was significantly reduced by the homogenization procedure. This is evidenced by the smaller standard deviations of the 40 sentences of the final speech material (filled bars, *after phase III*) in comparison to the standard deviation of the material employed in evaluation phase II (open bars, *before phase III*).

All standard deviations of the recognition rates calculated for the initial corpus material were compared with the standard deviations calculated for the 40 sentences per sentence type of the optimized corpus material. Using a Wilcoxon signed-rank test to compare the standard deviations before (median = 24.3) and after homogenization (median = 20.3), we found a significant decrease in the standard deviations for the optimized speech material ($z = -5.88$, $p < 0.001$, $r = -0.59$). The variance between sentences within one sentence type decreased by about 20% on average.

4. Properties of the final, optimized sentence corpus

The material of the OLACS corpus was tested for plausibility, as described above. The best recordings were then

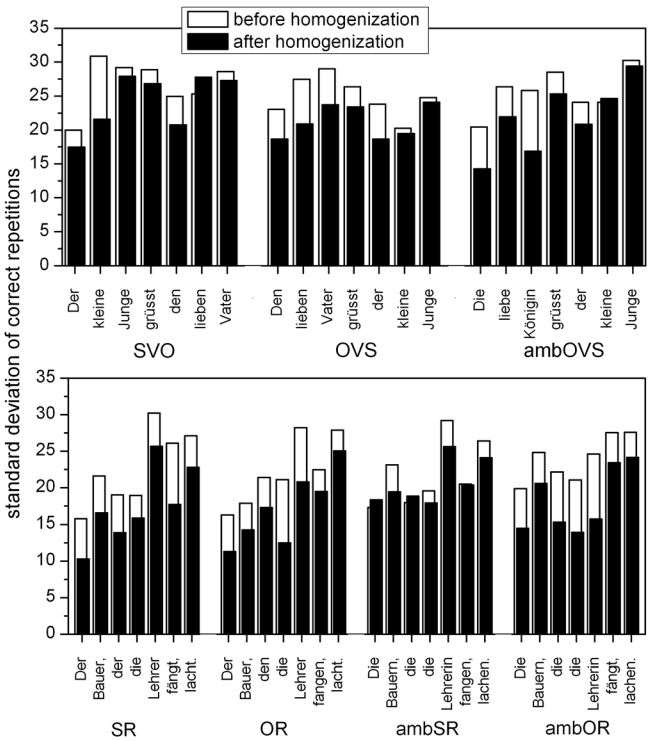

FIG. 3. Standard deviations of the correct repetitions after evaluation phases II and III at each word position when averaged across sentences for verb-second sentences (upper panel) and relative clause sentences (lower panel). Open bars depict the standard deviations for the whole material tested in phase II (before homogenization) and filled bars depict standard deviations for only the 40 sentences of each type, which now make up the material for speech intelligibility tests (after homogenization).

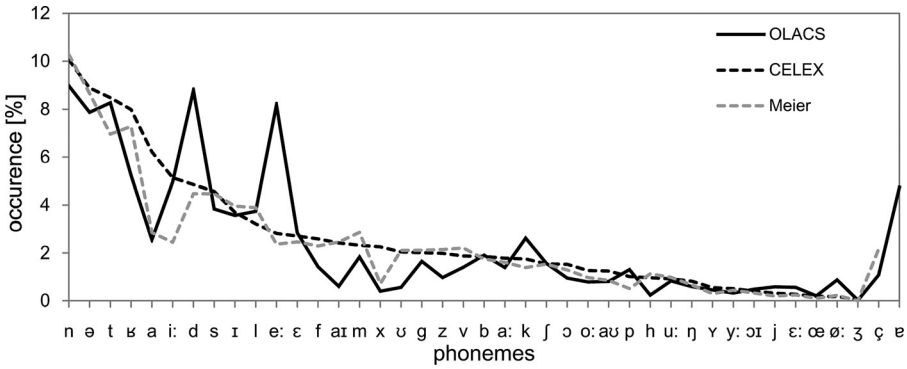

FIG. 4. Mean phoneme distribution of the OLACS and mean phoneme distribution in the German language according to the analysis by Hofmann *et al.* (2007) of the CELEX corpus and according to Meier (1967).

carefully selected for their naturalness of the speech, their speech rate, and certain prosodic characteristics. Afterwards, the homogenization of the intelligibility of the speech material took place in three consecutive stages. During the optimization procedure the sentence material was reduced from initially 680 sentences to 280 sentences with 40 sentences of each sentence type. This procedure led to the final, optimized sentence corpus for speech intelligibility measurements. General properties of said corpus are described below.

a. Linguistic measures. The final 280 sentences contain 102 nouns in different grammatical forms (male/female, singular/plural, nominative/accusative), 26 different transitive verbs (either as the verb in the sentence types 1–3 or in the relative clause; in third person singular or plural), 10 different intransitive verbs (as the verb of the matrix sentence in the second sentence category; in third person singular or plural), and 52 different adjectives (male/female, nominative/accusative, always singular). All sentences have between 11 and 13 syllables. The mean speech rate for verb-second sentences is 243 ± 24 syllables per minute and the mean speech rate for relative clause sentences is 206 ± 20 syllables per minute. This difference is presumably due to the longer pauses at the second comma in the relative clause sentences (see the acoustical analysis of the OLACS material in Carroll, 2013).

The phoneme distribution of the remaining 280 sentences was analyzed and compared to the mean phoneme distribution of the German language according to Meier (1967) and the phoneme distribution analysis of the CELEX corpus (Baayen *et al.*, 1995) by Hofmann *et al.* (2007). There were no significant differences, although four rather large deviations were observed between the phoneme distribution of the OLACS material and the CELEX corpus (Pearson's $r^2 = 0.838$ and $p < 0.001$, for each comparison; Fig. 4). For the phoneme /a/, the distribution found for our material follows the distribution postulated by Meier (1967), whereas the distribution found in the CELEX corpus diverges from both, the distribution in Meier and our distribution. The phonemes /d/ and /e/ are somewhat overrepresented in our corpus because of the prevalent use of *der* and *den* in our sentences. Last, our transcription deviates from the CELEX corpus and Meier (1967) in using the phoneme /ɐ/ for vocalized "er" in words such as *Metzger* "butcher" /mɛtsgɐ/. Meyer and CELEX transcribed those words with /əʁ/ (/mɛtsgəʁ/). We argue that the use of /ɐ/ gives a more realistic account of the pronunciation used by many standard German speakers and particularly by our speaker.

b. Speech intelligibility measures. The resulting mean sentence discrimination functions of the 40 sentences of each sentence type are depicted in Fig. 5. They were determined using the maximum likelihood estimator described in Eq. (2). Input for each fit consisted of the recognition rates of all 40 sentences per sentence type at the three SNRs each sentence was measured at during evaluation phase II and III. Thus, each fit is based on 120 data points.

The SR type showed the lowest SRT (corresponding to 50% recognition rate) and the OVS sentences showed the highest SRT (difference in SRT: 2.6 dB). Also, the ambiguous SR sentences showed a higher SRT than the respective non-ambiguous type. The ambiguous OVS sentences showed the same SRT as the SVO sentences but a shallower slope of the intelligibility function than all other sentence types. This shallower slope might be due to a slightly greater variance in the values within the individual ambOVS sentences.

In conclusion, the shapes of the sentence-type-specific discrimination functions differed considerably even though the underlying word material was selected to have similar intelligibility when presented without the context of the complete sentence. This is additional proof for the viability of OLACS and may be due to subject-before-object preferences of the listeners and inherent context effects of the sentence material.

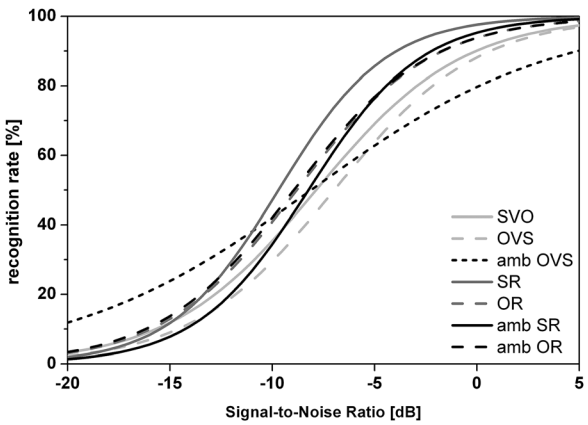

FIG. 5. Mean sentence discrimination functions (recognition rate [%] as a function of SNR ratio [dB]) for each class of sentences from Table I.

B. Verification of the test design: Effect of linguistic complexity on speech reception thresholds and response onset times

According to the design principles of OLACS, linguistic complexity is expected to have a consistent and systematic effect on test outcomes. To verify this, response latency and sentence intelligibility were compared between complex sentence types. As an additional test parameter to a quiet test condition, a masker (stationary or fluctuating noise at 65 dB SPL) was employed to test the differential effects of linguistic complexity with varied complexity of the acoustical conditions.

1. Participants

Twenty native German students (7 male and 13 female; mean age: 24 ± 2 yr) of various disciplines at the CvO University of Oldenburg participated in this study. All participants had otologically normal hearing, defined here as having a pure tone audiometry hearing threshold of 10 dB hearing level (HL) or better at the standard audiometric frequencies in the range between 125 and 8000 Hz. With the exception that for a maximum of two measured frequencies the hearing threshold was allowed to be at maximum 20 dB HL. Pure tone average (PTA) ranged from −5 to 10 dB. None of the listeners had previously participated in experiments in which speech intelligibility tests were used in any form. Conducting all experiments took about 4 hs (including breaks), evenly distributed across two sessions. The same investigator conducted all experiments.

2. Test procedures

a. Speech intelligibility measurements with OLACS. For the measurement of speech recognition with OLACS, listeners were presented with one spoken sentence at a time. The listeners' task was to repeat the sentence as accurately as possible. The investigator scored each correctly repeated word, obtaining a recognition rate for each sentence. Each word of the sentence had the same weighting, regardless of whether or not it was a key word (i.e., a semantically full lexeme).

Sentences were presented in quiet or against one of two different background noises in counterbalanced order. One test block used the same stationary speech shaped noise as was used in the evaluation procedure. Another test block used the icra4-250 noise, a speech-shaped noise with a female frequency spectrum and fluctuations of a single talker (original icra4 noise by Dreschler *et al.*, 2001, modified according to Wagener *et al.*, 2006, to a maximum pause length of 250 ms). Both noise types were presented at an unweighted root mean square value which was equivalent to 65 dB SPL.

The SNR (or the level of the speech in the quiet condition) changed adaptively throughout the measurement to determine the SNR/speech level at which 80% of the speech material was understood correctly (thus assessing the SRT_{80}) for each sentence type. The level of speech was increased if 80% or more of the presented words in a sentence were repeated inaccurately and decreased if more than 80% of the words were repeated correctly (see Brand and Kollmeier, 2002, procedure A1, with an estimated slope of 15% per dB for the discrimination function of the OLACS material and a different target intelligibility). Measuring at the 80% point of the intelligibility curve represents a more realistic listening situation than at the 50% point. Moreover, for our purpose listeners needed to get beyond the word recognition problem at the lexical level in order to encounter differences at the sentence level.[5]

Since the sentence material uses a limited vocabulary, and since it uses seven distinct and therefore easily learnable sentence types, there were some training effects to be expected in addition to the increasing familiarity with the procedure and the voice of the speaker. Therefore, each listener first completed one training list with 28 sentences at a fixed SNR of −2 dB in stationary noise, which contained four sentences of each type, in order to reduce the influence of training effects during the actual measurement. At the end of said training list, none of the participants made any further mistakes. After the training list, participants were presented with the three test blocks (quiet, stationary noise, and fluctuating noise). Each test block contained 140 sentences, such that 20 sentences of each sentence type were included. The order of the three test blocks and of the sentences in each block was randomized for each listener. The randomization across different sentence types was necessary in order to prevent the listeners from benefitting by learning the structure of the respective sentence type. The stimulus material was presented using an in-house MATLAB-based script that kept track of the responses and the SNR/speech level for each sentence type separately and changed the SNR/speech level independently for each sentence type in an interleaved manner. Thus one SRT_{80} could be obtained for each sentence type in each of the three test blocks. Measurement time for each block was about 30 to 35 min.

b. Response latency measurements. During the speech intelligibility measurements described above, trial recordings were made of each presented sentence (rerouted to a recording device) in one channel and the participant's response (recorded by microphone and routed to the same recording device) in the other channel of a stereo recording. After the measurements, response latencies (i.e., the time between stimulus offset and the onset of the listener's vocal response) were calculated for each trial after the actual measurements using a MATLAB script which detected voice on- and offsets. Though the script worked very reliably, we checked each calculated value afterwards to exclude errors due to coughing or other noises recorded in addition to the answer of the participant. The response onset times of the 11 participants for whom we had complete recordings were analyzed for all three listening conditions (quiet, stationary, and fluctuating noise). The last six trials of each sentence type presented to each listener were included in the analysis. This amounted to a total of 1376 sentences that were included in the analysis.

c. Cognitive tests. All participants were tested on a battery of cognitive tests in order to account for variation between subjects in the speech recognition study. At the

beginning of the first session they completed a Stroop task, a digit span (backward) test, and a word span (forward) test in random order.

The digit span test is part of the verbal HAWIE-R intelligence test (the revised German version of the Wechsler Adult Intelligence scale; Tewes, 1991). In the backward version used here the participant was presented aurally with a chain of digits and the task was to repeat the chain in reversed order. This test has been linked to a combination of memory and processing capacity and the ability to manipulate the content of working memory (e.g., Cheung and Kemper, 1992). The word span test is based on the same design but uses words (two and three syllables) instead of digits (see Schuchart, 2008). Since the word span is tested forward only and consists of semantically unrelated words, the measure is related primarily to verbal short-term memory storage, without need for manipulation.

In the Stroop test (Stroop, 1935), participants had to decide if the meaning of the word displayed on a monitor fit the color of a rectangle displayed above the word. The font color had to be ignored. We used the test paradigm described by Kim *et al.* (2005), in order to gauge participants' susceptibility to interference from distracting stimuli.

3. Statistical methods

The SRT_{80} for each sentence type in each listening condition (quiet, stationary noise, and fluctuating noise) was calculated as the median of the speech level/SNR of the last six sentences presented for each sentence type for each listener. Statistical differences between sentence types were all calculated with repeated-measures analyses of variance (ANOVAs) using SPSS 18. Significant effects were followed by pair-wise comparisons using *t*-tests. Bonferroni correction was used where appropriate. The influence of different cognitive measures and other person-level factors was assessed using factor analysis and calculating correlation coefficients.

III. RESULTS

A. Effect on speech intelligibility

Figure 6 depicts the mean SRT_{80} values for each sentence type in each listening condition (quiet, stationary noise, and fluctuating noise). Even though the effect of sentence type on SRT_{80} was comparatively small (i.e., in the order of 2 dB), some of the differences across sentence types were significant, revealing a systematic pattern: SR sentences produced the lowest (best) SRT_{80} in each listening condition. From the verb-second sentences, the ambOVS type produced the lowest SRT_{80} in quiet and in stationary noise, whereas in fluctuating noise the ambOVS sentences produced the highest SRT_{80}. Note that these findings stand in contrast to the fits plotted in Fig. 5, whereby ambOVS should show the worst SRT_{80} even in the stationary noise. It further verifies the hypotheses that the fit might underestimate the slope of the ambOVS sentences to some extent. But for the validity of the results of the adaptive measurements presented here, this slightly greater variance should have no bearing, since the results are quite reproducible across participants. However, it might have some

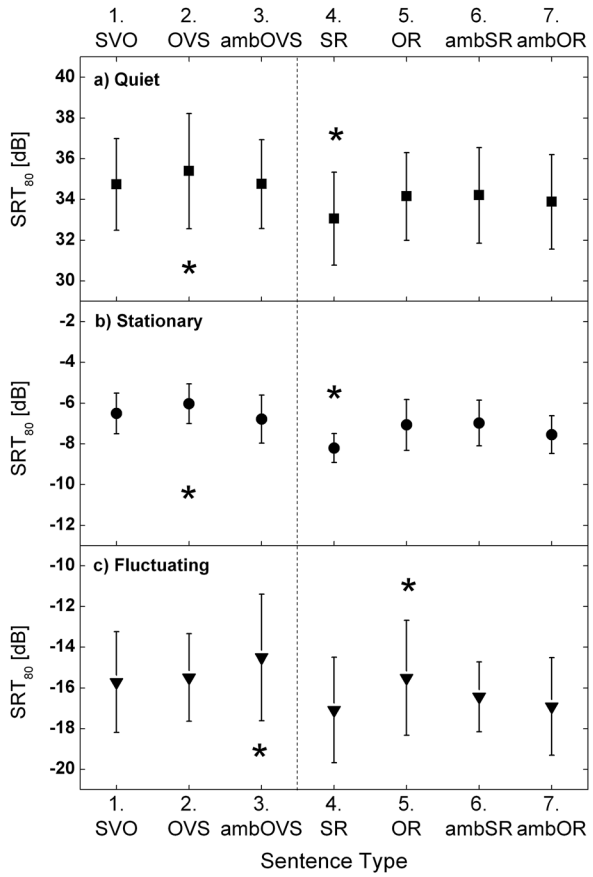

FIG. 6. Mean SRT_{80} and standard deviation across listeners for each of the seven sentence types and each noise condition. An asterisk standing above or below a specific sentence type indicates that this sentence type significantly differs from all the other sentence types in the respective category (verb-second and relative clauses). Note that for consistent comparison, the scale of the *y* axis always spans 10 dB.

influence on the test-retest reliability of the SRT of this sentence type. In stationary noise the inter-individual differences (not shown) were much smaller (data spread of about 3 dB), whereas in quiet and in fluctuating noise inter-individual differences exceeded 5 dB.

A two-way repeated measures ANOVA for the verb-second sentences (SVO, OVS, and ambOVS) and listening condition as inner subject factors revealed no significant main effect of sentence type. A significant effect of noise type [$F(2,18) = 9487.95$; $p < 0.001$] and a significant interaction between sentence type and listening condition [$F(4,16) = 4.208$ and $p = 0.004$] were found. The latter indicates that results for the three sentence types differed depending on the background noise employed.

This interaction effect warranted further investigation of the results using *post hoc* paired *t*-tests. For the SRT_{80} in quiet and in stationary noise the results of the *t*-test were nearly the same. There was a trend towards a difference between SVO and OVS sentences [$t(19) = -2.071$; $p = 0.052$; $t(19) = -1.908$; $p = 0.072$, respectively] and there was a significant difference between OVS and ambOVS sentences [$t(19) = 2.265$, $p = 0.035$ and $t(19) = -2.722$, $p = 0.014$]. Conversely, in fluctuating noise ambOVS sentences tended to differ from the other two sentence types [$t(19) = 2.087$, $p = 0.051$ and $t(19) = 2.066$, $p = 0.053$].

A two-way repeated-measures ANOVA for the relative clause sentences (SR, OR, ambSR, and ambOR) and listening condition revealed a significant effect of sentence type [$F(3,17) = 9.135$ and $p < 0.001$] and a significant effect of listening condition [$F(2,18) = 10.276$ and $p < 0.001$]. No interaction between listening condition and sentence type was found. Again, for the SRT_{80} in quiet and in stationary noise the results of the *t*-tests were nearly the same. SRT_{80} values of the SR sentences differed significantly from the SRT_{80} values of all other RT sentences [all $t(19) < -3.358$, $p < 0.005$ for the different combinations]. Conversely, in fluctuating noise OR sentences differed significantly from the other three sentence types [$t(19) > 2.358$, $p < 0.05$ for all three comparisons].

To better analyze the mean advantage of a simple sentence structure over a more complex sentence structure, the individual differences between the simplest sentences (SVO and SR) and each of the other two or three sentence types in the respective category was calculated. Figure 7 depicts the mean individual differences for each observed pair.

A one sample *t*-test was used to determine whether the difference between each pair differed significantly from zero, thus establishing whether the lower average SRT_{80} value for the simple sentence type was significant or not. In quiet and in stationary noise, participants benefited slightly but significantly (by approximately 0.5–1 dB) from the simple sentence structure in all but one case [SVO − ambOVS, $t(19) < -2.071$, $p < 0.05$ for all observed pairs]. In fluctuating noise only the SR − OR difference differed significantly from zero [$t(19) = -2.358$, $p = 0.029$] and there was a strong trend for the SVO – ambOVS pair [$t(19) = -2.087$, $p = 0.051$]. However, the benefit was larger in this condition (around 1 to 2 dB for each participant). Additionally, it should be noted that inter-individual variance was much smaller in quiet and in stationary noise than in fluctuating noise.

B. Effect on response latency

Figure 8 depicts the mean response onset times and the standard deviations averaged across listeners. There was a large variability in the recorded response onset time across listeners and listening conditions, which dominated the potential differences across sentence types and within each sentence type.

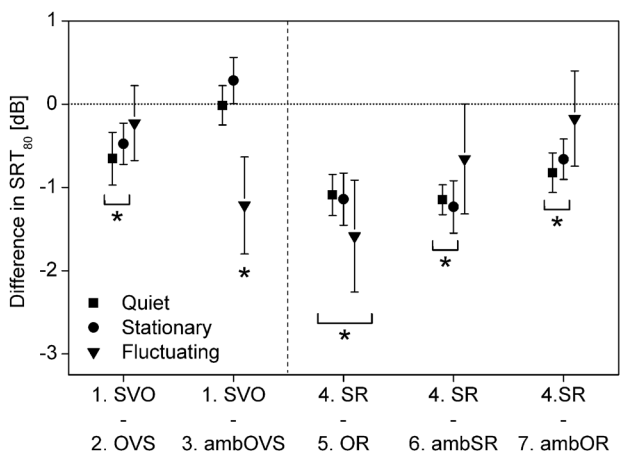

FIG. 7. Mean individual differences in SRT_{80} between sentence types and standard errors for the differences. Negative means indicate that the SRT_{80} for the simple sentence type was lower (better) than the SRT_{80} for the linguistically more complex sentence type. An asterisk denotes a group deviating significantly from zero.

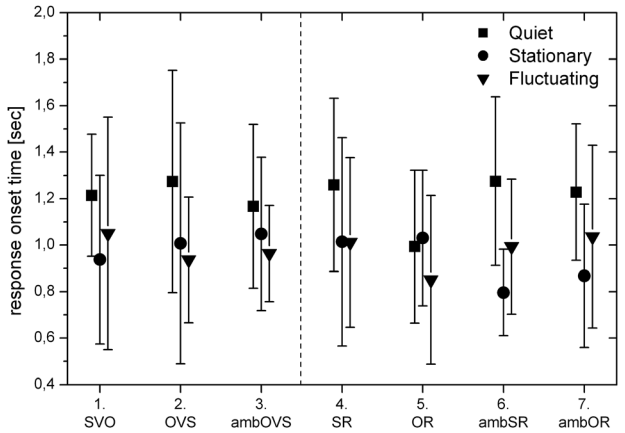

FIG. 8. Mean response latencies and standard deviations across listeners in each noise condition for each sentence type.

This large effect was analyzed first using a logarithmic transformation [$x_{1\log} = \log(x_1)$] on the data set for the statistical analysis in order to reduce the asymmetry of the underlying distribution of response latencies. Shapiro–Wilk tests for normal distribution for each sentence type in each listening condition were not significant after the transformation [for all $D(66) > 0.96$ and $p > 0.13$]. For verb-second sentences, a two-way repeated measures ANOVA with sentence type (SVO, OVS, and ambOVS) and noise condition (silence, stationary, and fluctuating noise) as inner subject factors revealed a significant effect of the listening condition [$F(2,18) = 15.641$, $p < 0.001$], and the pair-wise comparison showed that response onset times in quiet were significantly longer than in either noise condition ($p < 0.001$). No significant difference across these three sentence types was found, but a weak trend was observed [$F(2,18) = 4.723$, $p = 0.094$]. No significant interaction between listening condition and sentence type could be established. In addition, a one-way ANOVA with listeners as grouping factor (for between-subject comparison) revealed significant differences between listeners [$F(10,583) = 41.874$, $p < 0.001$].

For the relative clause sentences, a two-way repeated measures ANOVA with sentence type (SR, OR, ambSR, ambOR) and listening condition as inner subject factors revealed a significant effect of the noise condition [$F(2,18) = 11.563$, $p < 0.001$]. Again, the pair-wise comparison showed that response onset times in quiet were significantly longer than in the two noise conditions ($p < 0.001$). Also, there was a significant effect of sentence type [$F(3,17) = 9.006$; $p = 0.029$], which is mainly due to the response times observed in the OR condition. Again, no significant interaction between listening condition and sentence type could be established. Furthermore, a one-way ANOVA with listeners as grouping factor revealed significant differences between listeners [$F(10,781) = 104.172$, $p < 0.001$]. Hence, for verb-second as well as for relative clause sentences, there were systematic differences in response latency across sentence types. The large variance in response latency was due

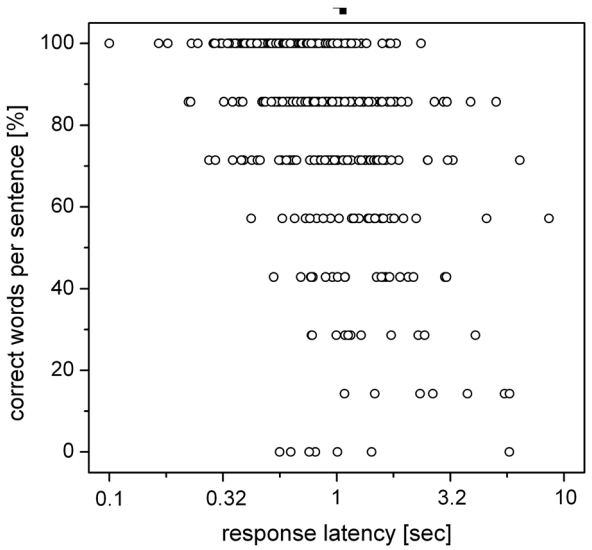

FIG. 9. Scatter plot for correlation in stationary noise between response latencies and the percentage of correctly repeated words per sentence. Note the logarithmic x axis, which accounts for the logarithmic transformation of the response latency data.

largely to differences between listeners, but also in smaller part due to differences between sentence types.

A high correlation was found between the response latency and the percentage of correctly repeated words per sentence [Pearson's $r^2 < -0.462$; $p < 0.001$ in all listening conditions, see Fig. 9 for scatter plot in stationary noise]. The scatter plot reveals that the spread of response latencies is higher the more words were repeated correctly. However, the lower boundary for the response latencies clearly increases with decreasing word recognition rate from about 100 ms to about 1 s. This holds across all cases except for those where none of the words were repeated correctly, here the response latencies decrease again. This might be due to the fact that if the listeners did not understand anything, there was no need to think longer.

Moreover, there was a small but highly significant correlation between the response latency and the speech level (or SNR) of the sentence [Pearson's $r^2 > 0.241$; $p < 0.001$ in all listening conditions]. The higher the speech level in the presented sentence, the faster the listeners were in giving their responses. However, the overall effect of the number of correctly repeated words had a stronger influence on response latency than did the level of speech.

Taken together, the results corroborate the test design in the sense that sentence type exerts a systematic effect on speech intelligibility: Even though the variability due to differences across listeners, listening conditions, percentage of correctly repeated words, and level of speech was very high, a small, but significant systematic effect of sentence type on response latencies could be established.

C. Cognitive measures

Cognitive processing was assessed in order to validate its effect on speech intelligibility using OLACS. An independent measure of cognitive processing was employed to evaluate if the individual differences in the OLACS results might be caused by differences in cognitive processing capacity rather than by differences in the sensory processing capabilities of the participants. Table II presents the results for the cognitive tests (word span, digit span, and

TABLE II. Results for the cognitive tests. Values for word and digit span are correct repetitions in percent; values for the Stroop test are mean reaction times in ms. Listeners are sorted into two groups by their results in the digit span. Age, sex, and pure tone average (PTA, mean of the audiogram values at 0.5, 1, 2, and 4 kHz in dB HL) are also provided.

Listener	Age	Sex	PTA	Forward word span	Backward digit span	Stroop
2	23	f	7.5	50	39.3	1003
19	30	f	−1.25	52.0	40.5	1060
1	23	f	10	38.8	51.2	1153
7	26	f	5.25	37.8	52.4	897
3	25	f	6.25	37.8	56.0	813
6	25	m	−2.5	55.1	56.0	857
17	24	f	−1.25	25.5	56.0	936
20	23	m	0	22.4	59.5	856
13	24	f	−2.5	40.8	71.4	923
4	26	m	5	36.7	73.8	1483
5	25	f	7.5	37.8	73.8	1147
14	21	f	6.25	48.0	85.7	794
12	26	m	−3.75	36.7	88.1	906
15	23	f	−5	27.6	88.1	1071
10	24	m	5	39.8	90.5	816
18	28	f	2.5	36.7	91.7	964
8	21	m	−5	52.0	92.9	886
16	22	f	3.75	51.0	92.9	889
9	23	f	3.75	38.8	94.0	883
11	23	m	−1.25	55.1	94.0	857
	24.3 ± 2.2 years		**2.0 ± 4.6 dB HL**	**41.0 ± 9.6 % corr**	**72.4 ± 19.3 % corr**	**960 ± 161 ms**

Stroop test) as well as the age, sex, and PTA of the participants. The word span produced lower values and a smaller spread than the digit span. The digit span task distinguished between two subgroups of participants in a statistically significant way (Mann–Whitney test for the two groups indicated by the horizontal line in Table II; U statistic $= 0.0$, $t = 36.0$, $p < 0.001$). However, the explanatory power of the digit span is limited, since this separation into two subgroups did not coincide with the individual performance in the other tests employed.

To analyze the impact of various parameters, we carried out a principal component analysis with the results of the three cognitive tests (word span, digit span, and Stroop test), PTA, age of the participant, and SRT_{80} for all sentence types in each listening condition. The principal component analysis was conducted with orthogonal rotation (varimax), and the Kaiser–Meyer–Olkin (KMO) measure verified the sampling adequacy for the analysis (KMO $= 0.83$). The analysis revealed two clearly distinguishable main components (after analysis of the scree plot), which together explained 47% of the variance in the data. Component 1 explained 26% of the variance and mainly contained the SRT_{80} in quiet and the PTA. Component 2 explained 21% of the variance and mainly contained the SRT_{80} in both types of noise and the outcomes of the individual and cognitive parameters entered into the analysis.

Figure 10 plots the loading of each variable for component 1 against the loading of the respective variable for component 2. A loading near zero shows that the variable has little or no influence on the respective component, whereas a loading of −1 or 1 indicates that the variable has a very high influence on the respective component. Loadings above 0.2 may be considered relevant. All measurements in quiet and the PTA had very high loadings on component 1 but small loadings on component 2. The measurements in noise (and here especially the measurements in fluctuating noise) and all individual and cognitive variables had relatively high loadings on component 2 but little loadings on component 1. The age of the participants (age span about 10 yr), and their performance in the digit span were of little importance in both components.

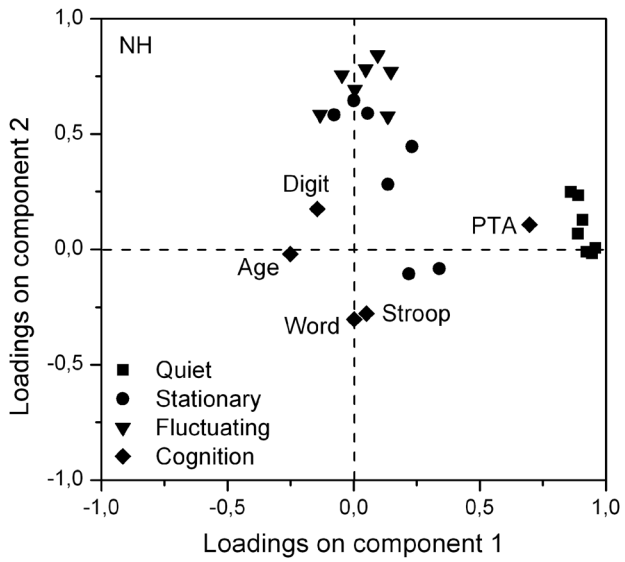

FIG. 10. Loading diagram of all variables included in the principal component analysis. The loadings for each variable on component 1 are plotted against the loadings on component 2. Each of the seven symbols in the three listening situations (quiet, squares; stationary noise, circles; fluctuating noise, triangles) represents one of the seven sentence types. The individual and cognitive variables (diamonds) are labeled accordingly.

The results of the principal component analysis were supported by the correlation coefficients for various variable pairs (two-tailed Pearson tests). The SRT_{80} of each sentence type was correlated with the PTA (see Table III) but no significant correlations were found between PTA and measurements in noise. This implies that for measurements in quiet the exact level of speech is more important than in noisy conditions. For the practical usability of the speech material this has only a small relevance since the standard deviation of the SRTs in quiet between sentences is small (about 2 dB) compared to the standard deviation of the PTA of the normal-hearing listeners partaking in this study (4.6 dB HL), and especially compared to the expectable standard deviations of PTAs in a potential clinical population In contrast, results from the Stroop test and both span tests correlated with some of the results for SRT_{80} in noise (see Table III) but not with the SRT_{80} in quiet.

Conversely, correlations between the individual differences in SRT (depicted in Fig. 7) and the cognitive variables revealed strong relationships between the differences in SRT in relative-clause sentences and both digit span and word span in quiet and in stationary noise (see Table III). Thus, someone with greater working memory capacity and/or a better ability to manipulate the stored content did not get worse SRT scores when presented with a more complex sentence. There were no statistically significant correlations between cognitive measures and individual differences in SRT in fluctuating noise. Also, the PTA did not correlate with the individual differences.

IV. DISCUSSION

A. Test construction and optimization

Although several approaches have been reported in the literature to relate linguistic complexity and speech perception (e.g., Tun *et al.*, 2010; Wingfield *et al.*, 2006), most of the test corpora used so far have either used fewer classes of test items to vary linguistic complexity, and/or have not controlled the acoustical information using a method as extensive as reported here. Hence, the present test is unique in its

TABLE III. Examples for significant correlations between different person level variables and SRTs.

Pair for correlation		Pearson's r^2	p
PTA	× SRT of SVO in quiet	0.463	0.031
PTA	× SRT of SR in quiet	0.490	0.028
word span	× SRT of OR in fluctuating noise	0.539	0.014
word span	× Difference of SRT SR-ambOR in stat.	0.560	0.010
digit span	× Difference of SRT SR-ambSR in quiet	−0.494	0.026

combination of controlling psycholinguistically and audiologically relevant parameters.

Besides acoustic parameters and the linguistic complexity considered so far, speech intelligibility, as measured in standard audiology, also depends on other non-acoustical factors. For instance, the predictability of the sentence is important (Boothroyd and Nittrouer, 1988; Bronkhorst *et al.*, 2002). Several tests specifically addressed the difference in speech intelligibility between predictable ("the witness took a solemn oath") and unpredictable sentences ("John hadn't discussed the oath," SPIN test by Kalikow *et al.*, 1977; Pichora-Fuller, 2008, for a comprehensive review). Also, in highly predictable everyday sentences (Plomp and Mimpen, 1979) so-called linguistic entropy (i.e., the information content of linguistic stimuli) seems to be influential regarding speech intelligibility (van Rooij *et al.*, 1991). In the study of van Rooij *et al.*, sentences that scored highly on a "what is the next letter" guessing procedure were easier to understand than sentences that scored low on the guessing task. In the current study, however, the predictability was chosen to be comparatively small, and constant across sentence types (as tested by the plausibility study), and thus in this aspect the material is more comparable to the matrix sentences as proposed by Hagerman (1982) and Wagener *et al.* (1999). The comparatively low variability of the SRT_{80} across sentence types observed here supports this assumption. One major cue to correctly interpreting our sentences was to understand who did what to whom. Since the probability of some persons doing something to others may influence perception and interpretation (i.e., the policeman catching the thief or vice versa), we ensured that the likelihood of the subject noun phrase doing something to someone was the same as the likelihood of the object noun phrase doing the same (cf. plausibility test). Note that in general the probability of occurrence is lower for non-canonical sentences (e.g., OVS, OR) than for canonical word orders (e.g., SVO, SR) for discourse and context-related reasons. Although this fact increased variability between structures, we were mostly interested in whether such structural predictability effects would reflect speech intelligibility measures. Future research will have to further investigate context-induced effects on speech intelligibility measures.

B. Verification of the test

The observed effect of linguistic complexity on speech intelligibility is in line with other studies that examined the effect of linguistic complexity on speech processing without necessarily using speech reception thresholds as an observable measure, although the effect was not as pronounced in the kind of task employed here. Wingfield *et al.* (2006) showed that, at least for younger listeners with normal hearing, the effect of linguistic complexity is small. In their study, comprehension scores changed from about 100% correct for subject relative sentences to about 96% correct for object relative sentences when sentences were presented at a normal speech rate. Furthermore, the effect size for these sentences with little semantic context was about the same as the effect found by Uslar *et al.* (2011) for young listeners with normal hearing for short, meaningful German sentences (Göttingen sentence test, GÖSA; Kollmeier and Wesselkamp, 1997; also denoted as "Plomp-type sentences," Plomp and Mimpen, 1979). The fact that there were only very small differences in SRT (about 1 dB difference between simple and complex sentences in the GÖSA and OLACS) supports Pichora-Fuller's (2008) conclusion that younger adults do not rely as much as older adults on semantic content.

The calculations of the SRT differences (Fig. 7) showed that all participants were taxed to the same degree by complex sentence structures in quiet and stationary noise. The results in fluctuating noise differed from those in quiet and in stationary noise: Both the inter-individual differences and the effect size were larger. This difference between listening conditions was expected, since understanding speech in fluctuating noise relies more heavily on more acoustically related processing abilities such as "listening in the dips" (see Bronkhorst, 2000, for a review). For listeners with normal hearing, dip-listening may decrease SRTs considerably, whereas for listeners with hearing impairment—depending on the measurement paradigm—the dip-listening benefit may not be as strong. For instance, even if the more favorable SNR hearing-impaired listeners need to correctly repeat a given proportion of the speech is accounted for, hearing-impaired listeners still show between 1 to 5 dB less benefit from the fluctuating masker (Bernstein and Grant, 2009). This is thought to be due to overall reduced audibility in terms of hearing threshold, distortions in the signal like deterioration of temporal and/or spectral resolution, and/or a decreased ability to use source-segregation cues (e.g., Rhebergen *et al.*, 2006; Bernstein and Grant, 2009). Also, listening in fluctuating noise may require a stronger cognitive contribution; this may increase the effect of linguistic complexity, since the capacities for resolving the complex sentence structure are needed for the perception of the speech in the noise background. It is therefore likely that the effect of linguistic complexity on speech intelligibility is even more pronounced when an interferer is used that produces more informational masking, such as other talkers (see also Mattys *et al.*, 2009; and a review by Kidd *et al.*, 2007).

It is especially noteworthy how the individual differences for SVO and ambiguous OVS sentences varied between the three listening conditions. In combination with the results of the evaluation phases I and II, in which fragments and complete sentences were presented, the diverging results in fluctuating noise stress the importance of distinguishing between speech intelligibility (i.e., the sensory-acoustical ability to perceive speech) and (language) understanding (i.e., the cognitive effort put into comprehending speech and putting it into the right context to ultimately correctly interpret an utterance). The OLACS material appears to provide a possible tool for examining this difference even on the level of the individual listener.

The findings of the response onset measures (Fig. 8) are in line with the findings of the speech intelligibility measurements in that they reveal systematic differences between sentence types. The correlation between response onset and the percentage of correctly repeated words per sentence indicates the consistency of both measures. Understanding more

of the sentence facilitates recall and thus results in faster response onset times. Additionally, online measurements like the reaction time studies by Carroll and Ruigendijk (2013) using the OLACS material, revealed significant effects of linguistic complexity during sentence comprehension. These online effects are measureable while the participants are still listening to the sentence before the observable end of the sentence and only in some cases even after the end of the sentence. These findings indicate that resolving syntactically more complex structures is more or less completed by the time the sentence is finished, at least in the simpler sentences contained in the OLACS.

Response latencies in noisy listening conditions may have been shorter than response latencies in quiet because the noise onset and offset gave the listener a better cue to direct their attention and confirm that the sentence was about to start and was finished, respectively. In contrast, in the quiet condition at a signal level where only about 80% of all words were understood correctly, there may not have been many cues as to the onset and offset of the sentence, possibly increasing the listener's uncertainty and thus prolonging the response onset time. An alternative explanation is that different processing strategies are used for listening in quiet and listening in noise: In a speech production study by Hanke *et al.* (2013) the response onset times were up to 80 ms shorter in noise than in quiet. This behavior might be linked to the higher processing load or memory cost incurred by cognition in noise. A faster speech onset could serve as a strategy to counteract more rapid fading of material from active memory.

As such, response latencies may have some clinically relevant implications, as they provide additional information about possible processing problems. But for clinical applications, the amount of data needed with this kind of measurement is likely to be too time consuming.

C. Considering individual subject data

Overall, the OLACS material indeed seems to induce a strong cognitive contribution, especially for complex sentences in adverse listening conditions, and more so than has been found for other materials (e.g., Akeroyd, 2008). The results of the factor analysis and the correlations indicate that, in quiet, speech reception thresholds correlate most strongly with the PTA, whereas in noise, cognitive factors gained in relative importance (see also Rönnberg *et al.*, 2010). Component 1 of the factor analysis can be interpreted as indicating general hearing ability in quiet, which is mainly characterized by the PTA. Component 2 represents hearing ability in noise as well as some form of cognitive contribution needed for the respective task. Still, both components together explain only about 50% of the variance of the data, with the cognitive component explaining about 20%. This is in line with findings of Rönnberg *et al.* (2010) and Akeroyd (2008). In his review, Akeroyd concluded that attention and working memory can explain at least part of the variance in speech intelligibility measurements. Note that we found that the benefit from canonical word order in relative clause sentences in quiet and in stationary noise depended mainly on working memory capacity, as indicated by the high correlations between calculated individual differences between sentence types and word span or digit span.

Conversely, in fluctuating noise there was no relation between cognitive measures and the benefit from canonical word order, but a rather large influence of cognitive measures on the overall performance in the speech intelligibility task, as indicated by the factor analysis. This provides additional support for the hypothesis that listening in fluctuating noise requires a higher cognitive contribution than, for instance, listening in quiet. Cognitive factors in this listening condition certainly help, as the overall performance was better for listeners with higher attention and/or working memory span. But attention and working memory do not seem to help any further in resolving the more complex sentences, since there was no relation between individual differences between sentence types and the cognitive measures employed here. In fluctuating noise, general auditory processing, like listening in the dips and discriminating between speech and background, appeared to be more important than the ability for fast and easy structural analysis of the sentence facilitated by working memory and attention.

Also note that in the study by Carroll and Ruigendijk (2013) with the same speech material, the influence of the same cognitive measures was larger than in our task. This may be because the authors ensured correct sentence comprehension by asking a who-did-what-to-whom question after each sentence they presented.

However, in the speech intelligibility measurements of the present study, comprehension and correct interpretation of the content does not seem necessary to repeat the sentence correctly, although it is certainly helpful, as evidenced by a high correlation between response latency and correctly repeated words. The more the listeners understood correctly, the faster they responded, thus indicating that greater understanding of the presented words facilitates recall. On the other hand, with OLACS, structural misinterpretation (i.e., taking the OVS sentence for an SVO sentence) may lead to reproduction mistakes in the answer of the participant (i.e., wrong casus). Thus, understanding the meaning might be more helpful in our sentences than it is for typical audiological material. For traditional speech intelligibility tests this implies the need for careful instructions in order to avoid instruction bias. Some subjects might have a strategy to repeat the words as (acoustically) heard without understanding the sentence, while others focus on the understanding and correct interpretation. Such instruction bias must be more pronounced in complex speech materials such as introduced in this manuscript and calls for research that examines the role of task instructions.

Future work with the material validated here for young normal listeners will employ listener groups that differ in age and hearing impairment in order to assess the differential effect of cognitive processing on speech intelligibility as a function of age and sensory impairment.

V. CONCLUSIONS

(1) The OLACS corpus established for speech intelligibility measurements was controlled for various linguistic factors such as careful selection of seven sentence types

that differentially vary in their prescribed linguistic complexity, word familiarity, the phoneme distribution, sentence length, and speech rate of the resulting test material.

(2) The OLACS corpus was audiologically controlled by selecting only those sentences and sentences that have a high homogeneity with respect to their item-specific speech intelligibility by selecting those sentences which were close to the specific mean of the measured SRT_{80} (a method used, for instance, by Kollmeier and Wesselkamp, 1997 or Ozimek *et al.*, 2009).

(3) The results of this study suggest the validity and applicability of the OLACS material as a tool to differentiate between acoustical factors and linguistic factors and their respective contribution in speech intelligibility measurements. Other studies (e.g., Carroll and Ruigendijk, 2013) furthermore show the suitability of the OLACS material in measurement paradigms more keyed to real understanding of the presented sentences.

(4) With OLACS we developed sentence material for which in at least one sentence type detrimental context effects can be found. These context effects warrant further studies.

(5) A principal component analysis and correlations revealed strong relationships between SRT_{80} measurements in noise and a subset of the cognitive tests, revealing the importance of working memory capacity (word span forward) and attention (Stroop test) in speech intelligibility tests with complex sentences. These findings support the initial premise that individual performance on tasks of varying linguistic complexity depends on individual cognitive capacities.

ACKNOWLEDGMENTS

The authors would like to thank Gary R. Kidd and the reviewers for their support with the manuscript. Supported by the Deutsche Forschungsgesellschaft (DFG), Grant Nos. KO 942/20-1, HA 2335/2-1, RU 1494/2-2, and BR 3668/1-2.

[1]Semantic roles denote semantic relations between the participants of an action. For example, the agent or actor role denotes the entity, which does or acts out the meaning indicated by the verb. This role is typically realized by a person. The patient role denotes the entity, which undergoes the action indicated by the verb. Since the patient role does not require any active engagement in the action denoted by the verb, it is often (but not always) realized as an inanimate object. For example, in the sentence "the man hits the thief," the man is the actor and the thief is the patient of the action "to hit." Here, the agent role is clearly allocated the grammatical function of subject, whereas the patient is allocated to the direct object.

[2]The argument structure of a verb refers to the selection of phrases that a given verb requires. For example, the argument structure of the verb "hug" consists of two arguments to build a grammatically correct sentence: a subject phrase denoting the hugger (e.g., the man) and an object phrase denoting the object or person to be hugged (e.g., the child, a tree). The verb "laugh," in contrast, only requires one argument, the laughing person, to construct a grammatically correct sentence ("the man laughs"). All additional information is optional. Some verbs allow for more than one argument structure. For example, the verb ask can either require a subject and two objects ("the man asks her a question") or a subject, an object, and an embedded sentence ("the man asked her how her day was"). All argument structures are assumed to be stored as part of the lexical entry in the mental lexicon.

[3]Most words and even phrases are repeated in several of the sentence types at different positions in the respective sentence For example, the word *Elefant*, "elephant," was both part of an SVO sentence (i.e., placed at the beginning of the sentence) and part of an OVS sentence (i.e., at the end of a sentence) as well as a part of sentences with a relative clause (either placed in the main sentence or in the relative clause). If *Elefant* is presented within a whole sentence of any type it is unclear whether a possible word intelligibility difference in the type of sentence is mainly caused by the respective acoustic representation or by the actual syntactic structure. By presenting the sentence fragments in evaluation phase I, we eliminated the effect of the sentence structure and therefore focused on acoustical cues.

[4]The slope for our optimized material is about 18%/dB and since the fragments are shorter, it is safe to assume that the slope for the fragments is shallower. For instance, the probabilistic model of Kollmeier (1990) (see also Wagener *et al.*, 2003) mathematically describes why intelligibility functions of sentences are steeper than the intelligibility functions for the single words which comprise the sentence.

[5]At the 50% correct level, not much cognitive capacity can go into integrating the heard words into a meaningful sentence, because the listeners do not recognize enough words to deduce the meaning of the sentence, whereas at the 80% correct level, cognitive effort put into integrating the words into a meaningful sentence might be worthwhile, because it could facilitate understanding. Moreover, a threshold at 80% intelligibility is less frustrating for the listener, as he can correctly repeat more items.

Akeroyd, M. A. (**2008**). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults," Int. J. Audiol. **47**(Suppl. 2), 53–71.

Altmann, G. T. M. (**1998**). "Ambiguity in sentence processing," Trends Cogn. Sci. **2**, 146–152

Baayen, R. H., Piepenbrock, R., and Gulikers, R. (**1995**). *The CELEX Lexical Database*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, http://www.mpi.nl/world/celex (Last viewed 11/27/12).

Bader, M., and Meng, M. (**1999**). "Subject-object ambiguities in German embedded clauses: An across-the-board comparison," J. Psycholinguist. Res. **28**(1), 121–143.

Benoît, C., Grice, M., and Hazan, V. (**1996**). "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," Speech Commun. **18**, 381–392.

Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U., and Wolff, C. (**2004**). "Language-independent methods for compiling monolingual lexical data," Lect. Notes Comput. Sci. **2945**, 217–228.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Bornkessel, I., Zysset, S., Friederici, A. D., von Cramon, D. Y., and Schlesewsky, M. (**2005**). "Who did what to whom? The neural basis of argument hierarchies during language comprehension," Neuroimage **26**, 221–233.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Bronkhorst, A. W. (**2000**). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acta Acoust. **86**, 117–128.

Bronkhorst, A. W., Brand, T., and Wagener, K. (**2002**). "Evaluation of context effects in sentence recognition," J. Acoust. Soc. Am. **111**, 2874–2886.

Caplan, D., and Waters, G. "Verbal working memory and sentence comprehension," Behav. Brain Sci. **22**, 77–94 (**1999**).

Carroll, R. (**2013**). *Effects of Syntactic Complexity and Prosody on Sentence Processing in Noise* (Shaker, Aachen).

Carroll, R., and Ruigendijk, E. (**2013**). "The effect of syntactic complexity on sentence processing in noise," J. Psycholinguist. Res. **42**(2), 139–159.

Cheung, H., and Kemper, S. (**1992**). "Competing complexity metrics and adults production of complex sentences," Appl. Psycholinguist. **13**, 53–76.
DIN EN 389-8 (**2004**). *Akustik—Standard-Bezugspegel für die Kalibrierung audiometrischer Geräte—Teil 8: Äquivalente Bezugs-Schwellenschalldruckpegel für reine Töne und circumaurale Kopfhörer (ISO 389-8:2004) [Acoustics—Reference Zero for the Calibration of Audiometric Equipment—Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones (ISO 389-8:2004), German version]*. European Committee for Standardization, DIN Deutsches Institut für Normung e.V. (Beuth, Berlin).
Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," Int. J. Audiol. **40**, 148–157.
Fanselow, G., Kliegl, R., and Schlesewsky, M. (**1999**). "Processing difficulty and principles of grammar," in *Constraints on Language: Aging, Grammar, and Memory*, edited by S. Kemper and R. Kliegl (Kluwer, Dordrecht), pp. 171–201.
Fanselow, G., Lenertová, D., and Weskott, T. (**2008**). "Studies on the acceptability of object movement to spec, CP," in *The Discourse Potential of Underspecified Structures*, edited by A. Steube (De Gruyter, Berlin), pp. 413–438.
Frazier, L., and Rayner, K. (**1982**). "Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences," Cogn. Psychol. **14**, 178–210.
Gordon, P. C., Hendrick, R., and Johnson, M. (**2001**). "Memory interference during language processing," J. Exp. Psychol. Learn. Mem. Cogn. **27**(6), 1411–1423.
Gorrell, P. (**2000**). "The subject-before-object preference in German clauses," in *German Sentence Processing*, edited by B. Hemforth and L. Konieczny (Kluwerpp, Dordrecht), pp. 25–63.
Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**, 79–87.
Hanke, M., Hamann, C., and Ruigendijk, E. (**2013**). "On the laws of attraction at cocktail parties: Babble noise influences the production of number agreement," Lang. Cognit. Processes (in press).
Hofmann, M. J., Stenneken, P., Conrad, M., and Jacobs, A. M. (**2007**). "Sublexical frequency measures for orthographic and phonological units in German," Behav. Res. Methods **39**(3), 620–629.
Just, M. A., and Carpenter P. A. (**1992**). "A capacity theory of comprehension: Individual differences in working memory," Psychol. Rev. **99**, 122–149.
Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351.
Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach N. I. (**2007**). "Informational masking," in *Springer Handbook of Auditory Research 29:Auditory Perception of Sound Sources*, edited by W. Yost (Springer, New York), pp. 143–190.
Kim, S., Kim, M., and Chun, M. M. (**2005**). "Concurrent working memory load can reduce distraction," Proc. Natl. Acad. Sci. U.S.A. **102**, 16524–16529.
Kollmeier, B. (**1990**). "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache" ["Measurement method, modelling, and improvment of intelligibility of speech"], Habilitation treatise, University of Göttingen, Fachbereich Physik, Göttingen.
Kollmeier, B., and Wesselkamp, M. (**1997**). "Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment," J. Acoust. Soc. Am. **102**, 2412–2421.
Lunner, T., Rudner, M., and Rönnberg, J. (**2009**). "Cognition and hearing aids," Scand. J. Psychol. **50**(5), 395–403.
MacPherson, A., and Akeroyd, M. A. (**2012**). "The rate of intelligibility change with level for continuous speech," *The Listening Talker*, Edinburgh, 2–3 May.
Mak, W. M., Vonk, W., and Schriefers, H. (**2002**). "The influence of animacy on relative clause processing," J. Mem. Lang. **47**, 50–68.
Mattys, S. L., Brooks, J., and Cooke, M. (**2009**). "Recognizing speech under a processing load: Dissociating energetic from informational factors," Cogn. Psychol. **59**, 203–243
May, C. P., Hasher, L., and Kane, M. J. (**1999**). "The role of interference in memory span," Mem. Cogn. **27**, 759–767
Meier, H. (**1967**). *Deutsche Sprachstatistik (German Language Statistics)* (Georg Olms, Hildesheim).
Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol. **41**, 329–335.
Ozimek, E., Kutzner, D., Sek, A., and Wicher, A. (**2009**). "Polish sentence tests for measuring the intelligibility of speech in interfering noise," Int. J. Audiol. **48**, 433–443.
Pichora-Fuller, M. K. (**2008**). "Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing," Int. J. Audiol. **47**(Suppl. 2), 144–154.
Pickett, J. M., and Pollack, I. (**1963**). "Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt," Lang. Speech **6**(3), 151–164.
Plomp, R. (**2002**). *The Intelligent Ear: On the Nature of Sound Perception* (Lawrence Erlbaum, Mahwah, NJ).
Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.
Reisberg, D. (**2007**). *Cognition: Exploring the Science of the Mind* (W. W. Norton, New York).
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.
Rodd, J. M., Davis, M. H., and Johnsrude, I. S. (**2005**). "The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity," Cerebral Cortex **15**(8), 1261–1269.
Rönnberg, J., Rudner, M., Lunner, T., and Zekveld, A. A. (**2010**). "When cognition kicks in: Working memory and speech understanding in noise," Noise Health **12**(49), 263–269.
Rudner, M., Rönnberg, J., and Lunner, T. (**2011**). "Working memory supports listening in noise for persons with hearing impairment," J. Acoust. Soc. Am. **22**, 156–167.
Schuchart, K. (**2008**). "Arbeitsgedächtnis und Lernstörungen: Differenzielle Analysen der Funktionstüchtigkeit des Arbeitsgedächtnisses bei Kindern mit Lernstörungen" ["Working memory and learning disabilities: Differential analyses of working memory functioning in children with learning disabilities"], doctoral dissertation, University of Göttingen.
Shapiro, L. P., Zurif, E., and Grimshaw, J. (**1987**). "Sentence processing and the mental representation of verbs," Cognition **27**(3), 219–246.
Stroop, J. R. (**1935**). "Studies of interference in serial verbal reactions," J. Exp. Psychol. **18**, 643–662.
Tewes, U. (**1991**). *Hamburg-Wechsler-Intelligenztest für Erwachsene—Revision 1991 (HAWIE-R)* (Huber, Bern).
Tun, P. A., Benichov, J., and Wingfield, A. (**2010**). "Response latencies in auditory sentence comprehension: Effects of linguistic versus perceptual challenge," Psychol. Aging **25**(3), 730–735.
Uslar, V., Ruigendijk, E., Hamann, C., Brand, T., and Kollmeier, B. (**2011**). "Sentence complexity effects in a German audiometric sentence intelligibility test: May we ignore psycholinguistics when testing speech in noise?," Int. J. Audiol. **50**, 621–631.
van Rooij, J. C., and Plomp, R. (**1991**). "The effect of linguistic entropy on speech perception in noise in young and elderly listeners," J. Acoust. Soc. Am. **90**, 2985–2991.
Vlaming, M. S., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B., Mohammadh, Y., and Houtgast, T. (**2011**). "HearCom: Hearing in the Communication Society," Acta Acust. Acust. **97**(2), 175–192.
Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
Wagener, K., Kühnel, V., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests" ["Development and evaluation of a speech intelligibility test for German I: Design of the Oldenburg sentence test"], Z. Audiologie **38**, 4–15.
Wagener, K. C., Brand, T., and Kollmeier, B. (**2006**). "The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners," Int. J. Audiol. **45**, 26–33.
Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., and Cox, L. C. (**2006**). "Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity," J. Am. Acad. Audiol. **17**, 487–497.
Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2011**). "Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response," Ear Hear **32**, 498–510.