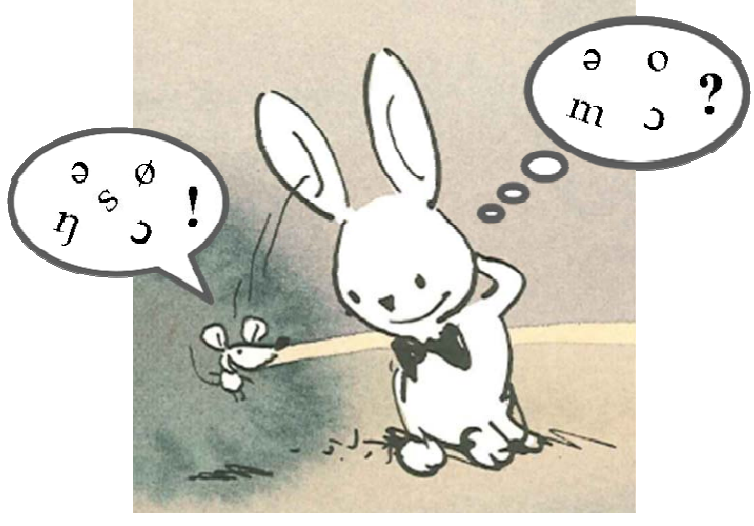
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 5

Jens Bo Nielsen

Assessment of speech intelligibility in background noise and reverberation



DTU Electrical Engineering
Department of Electrical Engineering

Assessment of speech intelligibility in background noise and reverberation

Ph.D. thesis by

Jens Bo Nielsen

Technical University of Denmark
2009

ISBN 978-87-92465-07-8
Printed in Denmark by Rosendahls - Schultz Grafisk a/s

Preface

This thesis is the result of my PhD studies at the Centre for Applied Hearing Research (CAHR) from March 2005 to August 2009. I have had professor Torsten Dau as my supervisor.

The main chapters of this thesis are based on three journal papers that can be read independently. However, the speech intelligibility test developed in chapter 3 is based on the test developed in chapter 2, so these two chapters are probably best read consecutively.

My interest for hearing research grew while I attended Torsten Dau's course *Auditory Signal Processing and Perception* when it was offered for the first time in the spring of 2004. I was on leave from my position at Nokia's R&D department in Copenhagen. Subsequently, I extended my leave in order to take an individual course at CAHR with the goal to develop a speech intelligibility test in Danish. The time frame was tight and the test was not finalized during the course. Fortunately, in January 2005, I received a PhD grant from the Oticon Foundation that permitted me to skip my job and continue my studies within speech intelligibility and perception.

I am very grateful that I was given the opportunity to be a PhD student at CAHR. I truly appreciate the scientific spirit and the supportive atmosphere that has made it possible for me to write this thesis.

Thank you to all colleagues at CAHR and Acoustic Technology for your help and assistance when I needed it.

A special thank you to Torsten for his tireless support and supervision. Thanks for pushing me towards higher standards than my own.

Thank you to the Oticon Foundation for funding this work.

Jens Bo Nielsen
Kgs. Lyngby, August 31st, 2009

Contents

Abstract

Reliable methods for assessing speech intelligibility are essential within hearing research, audiology, and related areas. Such methods can be used for obtaining a better understanding of how speech intelligibility is affected by, e.g., various environmental factors or different types of hearing impairment. In this thesis, two sentence-based tests for speech intelligibility in Danish were developed. The first test is the Conversational Language Understanding Evaluation (CLUE), which is based on the principles of the original American-English Hearing in Noise Test (HINT). The second test is a modified version of CLUE where the speech material and the scoring rules have been reconsidered. An extensive validation of the modified test was conducted with both normal-hearing and hearing-impaired listeners. The validation showed that the test produces reliable results for both groups of listeners. An important deviation between the two new tests and the original HINT is a new procedure used for equalizing the intelligibility of the speech material during the development process. This procedure produces more accurately equalized sentences than achieved with the original HINT procedure.

This study also investigates a fundamentally different method for assessing speech intelligibility. This method is based on the identification of the stop-consonant [t] in a short test-word. The method was originally developed in order to measure the impact of reverberation on speech intelligibility and, in particular, to measure whether the intelligibility of the test-word depends on the reverberation added to a surrounding speech carrier. It has been shown that the intelligibility of a reverberant test-word increases when the same amount of reverberation is also added to the carrier. In the literature, this observation has been interpreted as evidence of an *extrinsic compensation mechanism* for reverberation in the human auditory system. However, in the present study, it is shown that the listener's perception of the test-

word is not only related to the carrier reverberation but also to other of the carrier's acoustic-phonetic properties. The evidence of the extrinsic compensation mechanism is therefore questionable.

Overall, the results from the present study may contribute to the development of future speech intelligibility tests in Danish and other languages. The two developed sentence tests are expected to be useful for assessing speech intelligibility with Danish NH and HI listeners.

Dansk resumé

Bestemmelse af taleforståelighed i baggrundsstøj og efterklang

Pålidelige metoder til bestemmelse af taleforståelighed er af væsentlig betydning inden for høreforskning, audiologi og beslægtede områder. Sådanne metoder kan anvendes til at opnå en bedre forståelse af, hvordan taleforståeligheden påvirkes af f.eks. forskellige miljøfaktorer eller forskellige typer af hørenedsættelse. Denne afhandling beskriver udviklingen af to sætningsbaserede taleforståelighedsprøver på dansk. Den første prøve er Conversational Language Understanding Evaluation (CLUE), som er baseret på principperne i den oprindelige, amerikanske Hearing in Noise Test (HINT). Den anden prøve er en modificeret udgave af CLUE, hvor talematerialet og scoringsreglerne er blevet revurderet. En omfattende validering med både normalthørende og hørehæmmede lyttere er gennemført for den modificerede prøve. Valideringen viste, at prøven giver pålidelige resultater for begge grupper af lyttere. En vigtig afvigelse mellem de to nye høreprøver og den oprindelige HINT er den nye procedure, der anvendes under udviklingsprocessen til at udligne taleforståeligheden af sætningsmaterialet. Denne procedure giver mere præcist udlignede sætninger end den oprindelige HINT-procedure.

Denne afhandling undersøger også en fundamentalt anderledes metode til bestemmelse af taleforståelighed. Denne metode er baseret på identifikation af stopkonsonanten [t] i et kort test-ord. Metoden blev oprindeligt udviklet til måling af virkningen af efterklang på taleforståeligheden, og specielt om forståeligheden af test-ordet afhænger af efterklangen i en omgivende bæresætning. Det er blevet påvist, at forståeligheden af et test-ord med efterklang stiger, når den samme mængde efterklang også lægges på bæresætningen. I tidligere forskning er denne observation blevet tolket som et bevis på en *ydre kompensationsmekanisme* for efterklang i den

menneskelige hørelse. I den foreliggende undersøgelse er det derimod påvist, at lytterens opfattelse af test-ordet ikke kun afhænger af bæresætningens efterklang, men også af andre af sætningens akustisk-fonetiske egenskaber. Beviset for den ydre kompensationsmekanisme er derfor tvivlsomt.

Samlet set kan resultaterne fra denne afhandling formentlig bidrage til udviklingen af fremtidige taleforståelighedsprøver på dansk og på andre sprog. De to udviklede høreprøver forventes at være nyttige ved måling af taleforståelighed med danske normalthørende og hørehæmmede lyttere.

List of abbreviations

ANOVA	Analysis of variance
BRIR	Binaural room impulse response
CLUE	Conversational language understanding evaluation
HI	Hearing-impaired
HINT	Hearing in noise test
HL	Hearing level
NH	Normal-hearing
RMS	Root-mean-square
SI	Sentence intelligibility
SNR	Signal-to-noise ratio
SPL	Sound pressure level
SRT	Speech recognition threshold
SRT_N	Speech recognition threshold in noise
SRT_Q	Speech recognition threshold in quiet
STI	Speech transmission index
WI	Word intelligibility

Chapter 1

General introduction

Speech perception and, in particular, the assessment of speech intelligibility is a major field within hearing research and audiology. This is not surprising, since speech is the most important communication channel between humans. At the dinner table, in the classroom, or in the pub, speech is the main means of communication. Intelligible speech plays an important role, also in primarily visual media such as television and movies. Most speech communication takes place without the involved parties even giving the situation a thought, and the complicated process of encoding and decoding speech messages is generally taken for granted. The analysis of the process is facilitated if speech communication is regarded as a chain consisting of three stages: the talker (source), the transmission system, and the listener (receiver). All three components will affect the speech intelligibility, i.e., how much of the talker's intended message is understood by the listener. The talker affects the intelligibility by pronunciation, speech intensity, the complexity of the spoken message, etc. The transmission system affects the intelligibility by narrowing the bandwidth or distorting the signal in other ways. Reverberation in a room or the presence of background noise are common effects of the transmission system. The listener affects the intelligibility by his or her individual ability to decode the message. The auditory system, from the outer ear to the central auditory processing, is a key element influencing this ability, including effects of cognitive abilities, concentration, attention, etc.

Speech is robust. Despite the very diverse circumstances under which we communicate, the message normally gets through. Differences in pronunciation or speaking style, the presence of background noise or reverberation, the reduced bandwidth of a telephone line are just some of the factors that can severely alter a speech waveform,

yet leave it intelligible. The normal-functioning auditory system can process the incoming sound signal so effectively that the listener hardly notices how much the environment impacts the signal.

However, it is a common experience that speech communication is not flawless. Sometimes a word is misunderstood, sometimes whole sentences are incomprehensible. There can be innumerable reasons for this. The reverberation at a train station can be *too* strong for the listener to clearly understand the announcements. The background noise in a canteen can be *too* loud to make conversation possible, despite the ability of the auditory system to compensate for part of these effects. In particular, such conditions often represent a major challenge for people with a hearing impairment. While a moderate hearing loss may not affect speech intelligibility in quiet environments, the situation may be very different in an environment with background noise or reverberation. Even when the intensity of the speech signal is high, many hearing-impaired (HI) listeners experience difficulties in such situations. Speech intelligibility is, however, not always degraded by the transmission system. The early reflections in a room will normally enhance the intelligibility compared to a free field situation without any significant reflections. This effect is refined in carefully designed auditoria where a talker can be intelligible to hundreds of listeners without the need of amplification. Of particular interest in hearing research is the possibility to improve speech intelligibility by the use of hearing aids.

How listeners perform in a speech intelligibility task and how the performance is related to other characteristics of their hearing, e.g., their pure-tone audiogram, is a complex question. Although there is a correlation between the absolute hearing threshold for the pure-tones from 0.5 to 4 kHz and speech intelligibility in general, there are clear exceptions to this (e.g., Middelweerd et al., 1990). Listeners with almost normal audiograms can have severe problems with speech in noise, while listeners with a mild to moderate hearing loss might not have any problems as long as the reduced audibility of the signal has been compensated by amplification, e.g., by a hearing aid.

The complexity of the auditory system and its processing of speech makes it difficult to assess speech intelligibility reliably by indirect measures. This is the reason for the emergence of speech intelligibility tests that resemble an everyday speech

communication situation. A typical test consists of a series of speech stimuli that is presented together with a background noise. Such a test can be used for assessing the influence of both the transmission system and the listener on the speech intelligibility. Examples of such tests are the Speech Perception in Noise test (SPIN) (Kalikow et al., 1977), the Hagerman test (Hagerman, 1982), and the Hearing in Noise Test (HINT) (Nilsson et al., 1994). These tests are applicable for assessing the speech intelligibility problems that some listeners might have despite a normal or close-to-normal audiogram. An investigation related to the transmission system could be a measurement of how a specific hearing aid influences speech intelligibility. Effects of the acoustical environment, e.g., reverberation, can be investigated with a test setup in different physical locations or by adding reverberation to the speech material before it is presented to the listener over headphones.

A commonly used measure for the result of a speech intelligibility test is the *speech recognition threshold in noise* (SRT_N), which is the signal-to-noise ratio where 50 % of the presented speech material is correctly repeated by the listener. The Danish Hagerman test, DANTALE II, is based on the measurement of the SRT_N and for NH listeners the average result will be approximately -8.4 dB (Wagener et al., 2003). For some testing purposes, this value is too low and therefore a Danish speech test with a higher SRT_N has been of interest.

Chapter 2 describes the development of such a test, the *Conversational Language Understanding Evaluation* (CLUE). This is a speech intelligibility test with everyday sentences and it is intended for sentence-based scoring. In such a test, an approximately equal intelligibility of all sentences is important (MacLeod and Summerfield, 1990). This equalization is often done by a procedure based on objective scoring of the individual word intelligibilities (e.g., Soli and Wong, 2008). However, it is shown in this chapter that such a procedure does not lead to the intended equalization. Instead, the development process of CLUE introduces a new method for equalizing these intelligibilities. The method is based on a subjective assessment of the sentences done by NH listeners.

Although similar to the HINT, the CLUE test cannot be referenced as the “Danish HINT”. This and some other CLUE related issues led to the development of a Danish version of the HINT. Chapter 3 reports on the steps that were taken to create it. The

test is based on the CLUE speech material and it is validated with both NH and HI listeners. The NH and HI listeners were also retested after three weeks in order to investigate whether the test can be used more than once with the same listeners. A training effect was expected to affect the results of the retest. This effect can be split into a separate learning effect and memory effect, and the distribution between the two effects is investigated for a subgroup of the HI listeners.

In chapter 4, an alternative method for assessing the impact of reverberation on speech intelligibility is investigated. This method is based on the listeners' ability to detect a phonetic detail in a test-word instead of the overall ability to understand the speech material as in the previous tests. The listeners are asked to identify the stop-consonant [t] in the test-word when different amounts of reverberation are added to the word itself or to a surrounding speech carrier. When fewer identifications of the [t] are made, the speech intelligibility is assumed to be reduced. The method was originally presented in Watkins (2005c) where it was used to prove the existence of a *compensation mechanism* for reverberation in the auditory system. This mechanism was assumed to improve the speech intelligibility when the listener had had a short time to adapt to the reverberation. In chapter 4 some of Watkins' measurements are repeated and new results with additional carriers are obtained in order to investigate the validity of the test method and the evidence of the compensation mechanism.

Chapter 5 summarizes the main findings of the thesis and discusses implications for future research within speech intelligibility.

Chapter 2

Development of a Danish speech intelligibility test

This chapter is based on Nielsen and Dau (2009b)

Abstract

A Danish speech intelligibility test for assessing the speech recognition threshold in noise (SRT_N) has been developed. The test consists of 180 sentences distributed in 18 phonetically balanced lists. The sentences are based on an open word-set and represent everyday language. The sentences were equalized with respect to intelligibility to ensure uniform SRT_N assessments with all lists. In contrast to several previously developed tests such as the Hearing in Noise Test (HINT) where the equalization is based on scored (objective) measures of word intelligibility, the present test used an equalization method based on subjective assessments of the sentences. The new equalization method is shown to create lists with less variance between the SRT_Ns than the traditional method. The number of sentence levels included in the SRT_N calculation was also evaluated and differs from previous tests. The test was verified with 14 normal-hearing listeners; the overall SRT_N lies at a signal-to-noise ratio of -3.15 dB with a standard deviation of 1.0 dB. The list-SRT_Ns deviate less than 0.5 dB from the overall mean.

2.1 Introduction

Understanding speech is a fundamental human ability and listening to spoken language is probably the most important application of our hearing. Therefore, methods for a reliable assessment of speech perception capabilities are essential, particularly when hearing difficulties are suspected. An assessment must take into account that a hearing loss can affect speech intelligibility through at least two distinctly different sub-effects: (1) attenuation of all sounds entering the ear and (2) distortion of the perceived sounds (Plomp, 1978). Most hearing impairments are a combination of attenuation and distortion, but the distribution between the two parts varies from individual to individual. The effect of attenuation can be fully compensated by an increase in the overall sound pressure level, whereas the intelligibility loss due to distortion can only be compensated by an increase in the signal-to-noise ratio (SNR) (Middelweerd et al., 1990). Distortion can represent a considerable handicap in everyday situations because much speech communication takes place where the speech-to-noise ratio is low (Plomp, 1978). For a group of hearing-impaired listeners, the pure-tone thresholds at 500, 1000, 2000, and 4000 Hz are normally correlated with the speech intelligibility performance. However, a reliable prediction of the speech intelligibility cannot be made for the individual listener as only the attenuation part of the hearing impairment is measured directly in a tone audiogram. The audiogram is an inadequate method for predicting speech intelligibility, especially in noise, where the intelligibility is more affected by distortion than by attenuation (e.g., Glasberg and Moore, 1989; Middelweerd et al., 1990).

Various speech intelligibility tests that take distortion effects into account have been developed over the last decades. The earliest tests were based on short words presented in noise (e.g., Fairbanks, 1958; House et al., 1965), and the intelligibility score was calculated as the percentage of correctly repeated words. While including important features for the assessment of speech intelligibility, notably real speech stimuli and background noise, these tests have not been ideal for assessing a listener's ability to follow a natural conversation. The short, individually recorded words do not include many of the characteristics of natural speech, such as word transitions, reductions, contractions, temporal fluctuations and intonation (Nilsson et al., 1994).

Also, the listener's ability to exploit the redundancy as well as the semantic and syntactic cues in natural speech are not taken into account. Furthermore, word tests are not suited for more advanced testing and fitting of hearing aids, since the compression and noise-reduction algorithms do not take full effect with isolated single words (Nilsson et al., 1994). Examples of intelligibility tests using sentence-length stimuli are the Speech Perception in Noise test (SPIN) (Kalikow et al., 1977), the Hagerman-type test (Hagerman, 1982), and the Hearing in Noise Test (HINT) (Nilsson et al., 1994). The Hagerman-test was originally developed in Swedish and consists of five-word sentences constructed according to a fixed scheme: a name, a verb, a number, an adjective and finally a noun. The Hagerman-type test is also available in Danish (DANTALE II; Wagener et al., 2003). This test is suitable for extensive testing because the sentences are semantically unpredictable and difficult to memorize (Wagener et al., 2003). However, the sentences are also unnatural and nonsensical, and significant learning effects have been observed (Nilsson et al., 1994). Nilsson developed the HINT, which is a speech intelligibility test with natural sentences that comprise the pronunciation and content characteristics of conversational speech. The HINT mimics everyday speech communication and the test is sensitive to most of the speech perception problems encountered by the hearing impaired. The sentences are syntactically different and based on an open word set, which reduces the training effect compared to tests with a closed word set and a fixed sentence structure (Nilsson et al., 1994).

The outcome of a HINT measurement is usually the *speech recognition threshold in noise* (SRT_N), which is equal to the SNR at which the listener is able to correctly repeat 50% of the presented speech material. The use of the SRT_N effectively eliminates the risk of floor and ceiling effects where, respectively, 0% or 100% of the material is correctly identified. In the HINT, the SRT_N is measured using a sentence-based *adaptive* procedure where the SNR is decreased when the listener was able to repeat the whole sentence correctly, and increased when only part of the sentence was recognized.

The adaptive sequence is relatively short and the stimulus consists of 10 or 20 sentences taken from a set of pre-compiled lists (e.g., Vaillancourt et al., 2005; Hällgren et al., 2006). These lists must be of equal difficulty to ensure stable SRT_N

assessments. Additionally, it is a requirement for an adaptive procedure that the sentences within each list are of equal intelligibility. Otherwise, the SRT_N assessment will be unreliable and the test will be insensitive to small differences in the SRT_N between listeners or conditions (MacLeod and Summerfield, 1990). Equalization of the sentence intelligibilities is therefore an essential part of the test development process.

Sentences with equal overall root-mean-square (RMS) levels cannot be expected to be equally intelligible in noise, since word familiarity, short-term level variations, intonation, etc. will cause deviations (Nilsson et al., 1994). Therefore, a two-step process has been employed in the equalization of sentence intelligibility in several sentence tests (e.g., Plomp and Mimpen, 1979; Nilsson et al., 1994; Vaillancourt et al., 2005; Wong and Soli, 2005; Hällgren et al., 2006). In this process, the first step is to determine the intelligibility of all sentences that are candidates for the test. The sentences are presented to a number of listeners at various SNRs and their responses are recorded. In the second step, the intelligibility variations found in step one are compensated by an adjustment of the RMS levels of the individual sentences. The RMS level of sentences with low intelligibility is raised and that of sentences with high intelligibility is lowered. This adjustment exploits that the intelligibility of a sentence in noise is very sensitive to the SNR, thereby making it possible to compensate for intelligibility deviations by manipulating the SNR.

In step one of the equalization process, a considerable number of listeners is needed to obtain reasonably precise estimates of the sentence intelligibilities. In an attempt to increase the efficiency of the equalization procedure, previous HINT development projects have employed *word* scoring, noting the number of correctly repeated words in each sentence, instead of *sentence* scoring, noting only whether the whole sentence was correctly repeated or not. Compared to sentence scoring, word scoring increases the amount of collected data significantly. However, the use of word scoring in the equalization process has a severe side effect. The RMS adjustment of the sentences will no longer be based on the *sentence* intelligibility (SI), but on the average *word* intelligibility (WI) of each sentence. The WI for a given sentence is calculated as the number of correctly repeated words divided by the total number of words. In the studies on HINT (e.g., Nilsson et al., 1994; Vaillancourt et al., 2005;

Hällgren et al., 2006), it was assumed that sentences with equal WIs also have equal SIs, hence allowing an equalization of the SI of a corpus of sentences by ensuring that all sentences have the same WI. However, sentences with equal WI can indeed have significantly different SIs (see appendix A for examples of sentences with equal WI, but different SIs of 24% to 59%). This lack of proportionality between the SI and the WI is caused by the fact that the two entities have a different probabilistic relationship to the sentence. The SI is an "AND"-combination of what the listener repeats: all parts must be correctly identified in order to give the sentence a positive SI score. The WI is an "OR"-combination of what the listener correctly repeats: each individually identified word contributes positively to the WI score. The deviation between the SI and the WI will be particularly high when the distribution of the WI between the individual words in the sentence is very uneven. For example, if one word has a low intelligibility, the SI will also be low because it is seldom that all words in the sentence are understood. However, the words with high intelligibility will lead to a relatively high average WI.

It is essential that the SIs of the test developed here are equalized since the adaptive test procedure is based on sentence and not word scoring. This goal cannot be achieved with the use of word scoring in the equalization process, and the use of sentence scoring is extremely time consuming due to the large number of sentence presentations that are needed to achieve data of sufficient validity. To solve this dilemma, an equalization procedure based on a "just-follow-conversation" method (Kollmeier and Wesselkamp, 1997) was developed in the present study. The method builds on experiments where listeners are requested to adjust the individual SNRs of a number of sentences until the sentences are perceived as being equally intelligible. This method does not involve intelligibility tests with explicit sentence or word scoring. Instead, it involves subjective assessments of the sentences done by a group of listeners. The experiments of Kollmeier and Wesselkamp (1997) showed a correlation of 0.78 between the SRT_N found with the subjective method and previously conducted scoring experiments with the same sentences.

In the present study, the new equalization procedure was employed in the development of a speech intelligibility test for Danish. The objective was to develop a test with a minimal within-subject, between-list variation in the speech intelligibility

assessments, without reducing the sensitivity to between-subject variability. Since the developed test deviates in some methodological aspects from the HINT, it is named "Conversational Language Understanding Evaluation" (CLUE).

2.2 Methods

2.2.1 Sentence material

General selection criteria

The purpose of the present speech test is to evaluate a listener's ability to follow everyday conversational language. The sentence material for the test was created specifically with this purpose in mind. A set of "objective" and a set of "subjective" selection criteria were considered for the sentence material. The criteria are partly based on the criteria described in Versfeld et al. (2000).

The objective criteria are as follows: 1) The number of words in each sentence is 5. 2) The number of syllables in each sentence is 8-9. 3) Words do not contain more than 4 syllables. 4) Each sentence contains a verb. 5) The sentences are grammatically correct. 6) The sentences do not contain proper names. 7) Proverbs, exclamations and questions are not allowed.

The subjective criteria are as follows: The sentences should represent conversational speech and should be 1) neutral, 2) meaningful, 3) natural, and not be 4) (too) redundant, 5) illogical, 6) out of a context, 7) characterized by bad sentiments, or 8) humorous.

The Danish Society for Language and Literature ("Det Danske Sprog- og Litteraturselskab") has created a database named *Korpus 2000*, which contains sentences from Danish newspapers, magazines, books etc. from the period 1998 to 2002. There are about 28 million words in the database, corresponding to more than two million sentences. The sentences were tested against the criteria for inclusion in the test. The "objective criteria" (except for grammatical correctness) were implemented as MATLAB scripts, and 4075 five-word sentences were extracted from the database. A manual examination with respect to the "subjective criteria" and grammar showed, however, that most of the sentences were not usable. Many were context dependent

while others were characterized by (very) bad sentiments, e.g., “Vores liv ligger i ruiner” (our life lies in ruins) and “Der brændte bål i gaderne” (bonfires were burning in the streets). It was concluded that it would not be possible to extract enough usable sentences directly from Korpus 2000. A new corpus of sentences that fulfilled the listed objective and subjective criteria was written based on some of the five-word sentences extracted from the database. The sentences were then checked for naturalness and reoccurrences of words by the first author, and some were rewritten or discarded.

Selection of the talker

An initial recording of 25 sentences was made with 11 different talkers, four women and seven men. Their ages were 24-63 years. Five professional acousticians/audiologists reviewed the recordings of all 11 talkers. The objective was to find a talker with a natural pronunciation, close to the conversational speech that most Danes encounter in their everyday life. Seven talkers were rejected for various reasons, leaving four talkers still in consideration. The four recordings were assessed by two professional phoneticians. Based on their feedback, a 38 year old male talker with a background in phonetics, but with no previous experience in speech recordings, was selected.

Recording, editing and transcription

The sentences were recorded in a double-walled sound-proof booth directly to a PC using a high-quality 24-bit sound card (RME DIGI96/8 PAD) and a sampling frequency of 44.1 kHz. A 1-inch B&K condenser microphone (type 4179) with preamplifier (type 2660) was used to produce a recording with a low background noise level. The microphone was placed at a distance of approximately 30 cm from the mouth of the talker, symmetrically in the horizontal plane and at an angle of approximately 45^o in the vertical plane. The recorded sentences were digitally high-pass filtered at a cut-off frequency of 50 Hz and split into individual waveforms. The recordings were all adjusted to an average RMS level of -26 dB (re: max. digital output) with maximum peak levels of approximately -5 dB. This allowed

headroom for moderate level adjustments during the equalization process without the risk of clipping. The waveforms were stored as wav-files. Based on the sound files, the sentences were transcribed in the International Phonetic Alphabet (IPA). The transcription was done by a masters student in phonetics at the University of Copenhagen, who was highly experienced in transcribing spoken Danish.

The background noise

A speech-shaped background noise was created to match the long-term frequency spectrum of the sentence material. Speech-shaped noise maximizes the slope of the psychometric function, hence increasing the accuracy of the SRT_N determination (Prosser et al., 1991). The speech-shaped noise will also lead, on average, to similar SNRs across frequencies, hence not intentionally favoring some speech frequencies over others. However, some variation in the SNR will occur across the frequency spectrum due to the varying duration of different speech sounds. For example, for short and intense high-frequency consonants, the mean level of these (which by definition is the level in the speech-shaped noise) is relatively low due to their short duration, but when actually present in the speech signal, they will have a level well above the mean.

The noise was created using a superimposing approach (Wagener et al., 2003). The sentence sound files were concatenated in random order and stored as an initial noise file. The files were then randomized in a new order and added to the noise file. The final noise file was the result of 150 superpositions. The noise had only little amplitude fluctuation and a frequency spectrum that matched the long-term spectrum of the sentences. The RMS level of the noise was adjusted to the same level as the sentences.

2.2.2 Equalization of sentence intelligibility

Subjects

18 listeners (11 male, 7 female) participated in the equalization of the sentence intelligibility. Their ages were between 20 and 25 years with a mean of 22.8. Before participation, their audiograms were measured. All listeners had hearing thresholds

of 15 dB HL or better in the range 0.125 to 8 kHz. All listeners were native Danish speakers and students at the Technical University of Denmark. They were paid on an hourly basis for their participation. All experiments in this study were approved by the ethics committee of Copenhagen County.

Stimuli

The stimuli consisted of a corpus of 322 five-word sentences and the corresponding speech-shaped noise. The noise was turned off between sentences in order to avoid exposure to a constant noise. The noise onset was 1 s before the sentence start and the offset 600 ms after the end of the sentence. The noise was ramped on and off by a squared sine function with a ramp duration of 400 ms. The onset of the noise 1 s prior to the speech is believed not to create unintended onset effects and this timing has been used when determining normative data for the HINT in various languages.

Apparatus and procedure

The experiment took place in a sound-proof booth and stimuli were presented over Sennheiser HD580 headphones. The noise level was fixed at 65 dB SPL, whereas the speech level varied according to the listener's response. The experiment was controlled by a PC and a MATLAB application written for this specific purpose. The application presented the stimuli via the PC sound card to the headphones. An instructor was present in the booth for initial instruction and a short training session, but not during the experiment itself.

Experimental design

The equalization procedure of this study was fundamentally different from previous HINT studies. It was not attempted to objectively assess (score) the sentence or word intelligibility. Instead, the equalization was based on subjective judgements by the listeners. After each sentence presentation, the listener had the option to press one of three buttons: "difficult", "easy" or "ok". The listener was given written and oral instruction to press the buttons according to the following rules (translated from Danish):

- press “difficult”, if you did not understand the *whole* sentence
- press “ok”, if you were *just* able to understand the sentence
- press “easy”, if it was relatively *easy* to understand the sentence

The listeners were also instructed to create their own subjective criterion for how a sentence should sound to be “just understandable”, and only to press “ok”, when a sentence fulfilled this criterion. A test session with 12 sentences was run to help the listeners create the criterion before the actual equalization experiment started.

The sentences were presented in random order in a number of sequences. The first sequence consisted of all sentences, and the subsequent sequences consisted of all the sentences that had been judged “difficult” or “easy” in the previous one. A press of the “ok” button excluded the corresponding sentence from further presentation and stored the history of presentation levels for the sentence. The initial presentation level for all sentences was 63 dB SPL, corresponding to an SNR of -2 dB, the SNR at which listeners, on average, were expected to perceive the sentences as “just understandable”. A press of “difficult” raised the sentence level by 2 dB; a press of “easy” lowered the level by 2 dB. This adjusted level was then used for the presentation of the same sentence in the following sequence. After one reversal, for example, when a sentence was judged “easy” in one sequence and “difficult” in the next, the step size of the level adjustment was halved to 1 dB. At the second reversal, the step size was reduced to 0.5 dB, where it then remained.

The equalization process was split into two experimental series. 10 listeners participated in the first series and determined an RMS level for each sentence that subjectively equalized the intelligibility. All sentences were then adjusted with the mean adjustments done in this series and a second experimental series with 8 new listeners was conducted.

Outcome of the equalization

The number of sentence presentations to each listener in the first series ranged from 686 to 1456, with an average of 994, corresponding to 3.1 presentations per sentence. The mean adjustments of the sentences are shown in the left panel of

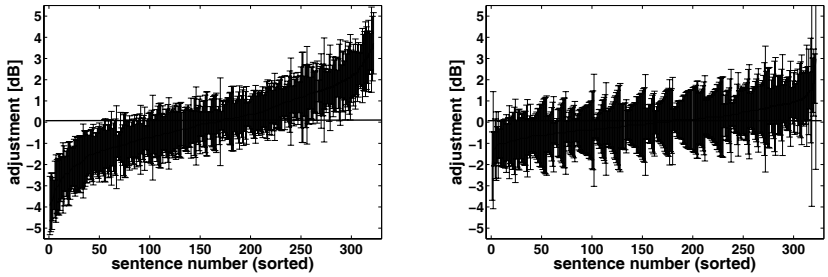


Figure 2.1: Mean adjustments of the sentence levels in the equalization process in the first experimental series (left) and in the second series (right). The adjustments are normalized to an overall adjustment of 0 dB. Error bars show ± 1 standard deviation. The 322 sentences are sorted with respect to the mean adjustment in each experiment. The adjustments in the first experiment are in the range -3.9 to 3.7 dB.

Fig. 2.1, normalized to an overall adjustment of 0 dB. The adjustments to produce "just understandable" sentences lie between -3.9 dB and 3.7 dB (averaged across listeners). The standard deviations lie between 0.44 and 2.45 dB.

In the second series, the number of presentations to each listener ranged from 701 to 1442; the average was 986, corresponding to 3.1 presentations per sentence as in the first series. The mean adjustments can be seen in the right panel of Fig. 2.1. The adjustments in the second series are, on average, smaller than in the first series, indicating that the level adjustments obtained in the first series had a positive effect on equalizing the intelligibility of the sentences. However, the adjustments obtained in the second series still deviate significantly from the "baseline" at 0 dB for a considerable number of sentences. Hence, the level adjustments determined in the second series were also imposed on the sentence files. The final sentence levels are thus based on the adjustments done in both series.

The level adjustments of the sentences were the immediate result of the equalization process, but the outcome also led to the omission of several sentences. Sentences with a total level adjustment of more than ± 3 dB were omitted in order to avoid obvious level differences in the final test. Adjustments of more than ± 1 dB were not allowed in the second series because a large adjustment might just be a statistical coincidence and there would be no third series to reveal this. Sentences were also omitted when a listener, during the equalization process, required an SNR of 4 dB or

more to comprehend the sentence. Such a high SNR implies an intelligibility flaw in the sentence. A minor, fixed adjustment was done to all sentences to ensure that the average RMS level was -26 dB (re: max. digital output).

2.2.3 List creation

The final sentence lists were created to be as phonetically balanced as possible. The sound inventory for the transcribed sentences consisted of 28 vowels (17 short and 11 long) and 20 consonants. The Danish "stød" (a short glottal stop) and syllabic consonants were also transcribed and regarded as phones to be balanced. The overall phonetic distribution for all sentences was determined and a trial-and-error procedure distributed the sentences among the lists in order to hit this distribution as closely as possible for each list. 20 lists with 10 sentences each were created.

2.2.4 List verification

The main purpose of the test verification was to document that similar SRT_Ns are obtained with the different sentence lists. The overall SRT_N for the test and its standard deviation were also determined.

Subjects

The verification of the 20 test lists involved 14 (7 male, 7 female) native Danish speaking listeners between 19 and 32 years (mean 22.9). They all had hearing thresholds of 15 dB HL or better from 0.125 to 8 kHz.

Apparatus and procedure

The experiment took place in a sound-proof booth and the stimuli were presented over Sennheiser HD580 headphones. The noise level was fixed at 65 dB SPL, whereas the speech level varied according to the adaptive test procedure of Nilsson et al. (1994). The noise onset and offset were controlled in the same manner as in the equalization procedure. A test leader was present during the whole experiment; he ran the test session using a tailor-made PC-application. The 20 test lists were presented to the

listeners in a random order determined by the PC. Before the actual test lists, 3 training lists were run. This allowed the listeners to get used to the test procedure and the influence of training effects was reduced.

The listeners were asked to repeat all words in the sentences as precisely as possible, but also encouraged to guess the words that they did not hear. The adaptive presentation procedure was as follows: The first sentence was presented repeatedly, starting at -8 dB SNR and increasing in 2 dB steps until the listener repeated the sentence correctly. The level of the remaining sentences was lowered by 2 dB after a correct repetition of the previous sentence, and raised by 2 dB after an incorrect repetition.

The PC application for running the test controlled the playback of the speech signal and the background noise and adapted the levels according to the test procedure. The test leader scored the sentences by pressing on-screen buttons according to the listener's response. The application created a data log of all sentence presentation levels during the test run.

Response variations

The following general variations of listener responses were accepted during the verification: 1) Change in verb tense, 2) change in article and 3) change between singular and plural nouns. Sentences were also considered correct if a word was added to the actual sentence. For example, "Han lagde tasken på bordet" (he put the bag on the table) was accepted in the form "Han lagde tasken op på bordet" (he put the bag up on the table). The following specific alternatives were also accepted: De/vi (they/we), hun/han (he/she), and min/din (my/your). In some cases these alternatives were mentioned spontaneously by the listeners.

2.3 Results

2.3.1 Calculation of the SRT_N

An analysis was conducted in order to determine the number of sentence levels to include in the SRT_N calculation. The difference between the presentation level of

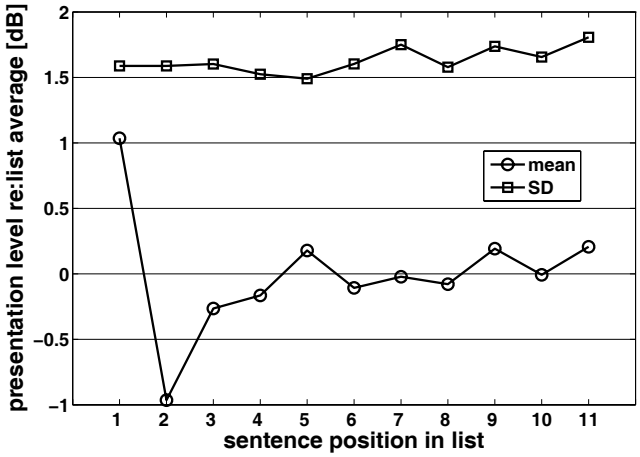


Figure 2.2: The presentation level mean and standard deviation across all lists and all listeners as function of the sentence position in the list. The presentation level is given relative to the average level of each sentence list.

each sentence in the verification test (n = 14 listeners · 20 lists · 11 levels = 3080) and the average level for each list (n = 14 listeners · 20 lists = 280) was computed. (Presentation level 11 results from the response to sentence 10, although the eleventh sentence does not exist.) Fig. 2.2 shows the mean of these level differences (circles) as a function of the position in the list. From the fourth sentence on, the presentation level has stabilized around the average; the level for sentence 4 is slightly closer to the average than levels 5, 9 and 11. The standard deviation (squares) has a minimum for sentences 4 and 5, but varies very little as a function of the sentence position. The result that sentence 4 is closer to the reference level than several of the following sentences was also found in the studies of Nilsson et al. (1994) and Hällgren et al. (2006). As a result, it was decided here to include the last eight levels (4 to 11) in the SRT_N calculation. This is a change in the calculation compared to Nilsson et al. (1994), who only included the levels of sentences 5 to 11.

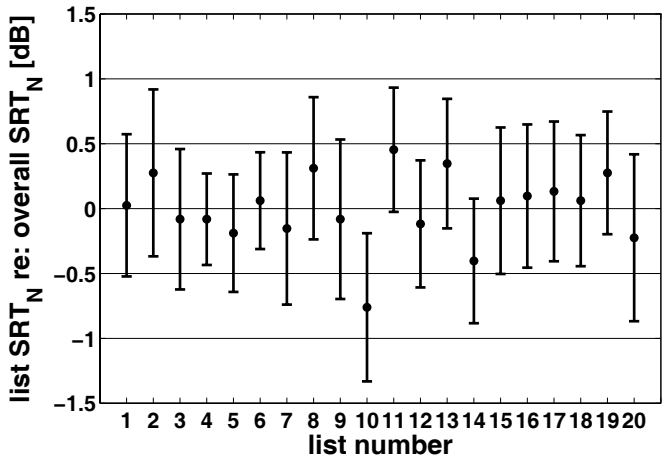


Figure 2.3: Mean list-SRT$_N$s deviations with 95% confidence intervals. When the 0-line is within the interval, the list-SRT$_N$ does not deviate significantly from the overall SRT$_N$.

2.3.2 List verification result

The following results all relate to SRT$_N$ calculations based on sentence levels 4 to 11 in each list. For each of the 20 lists, an estimated list-SRT$_N$ was calculated as the average SRT$_N$ across listeners. In Fig. 2.3, these values are plotted relative to the overall SRT$_N$, representing the mean SRT$_N$ across all lists and listeners. The 95% confidence intervals of the estimates are also shown, i.e., the interval around the mean that, with a likelihood of 95%, contains the "true" list-SRT$_N$. The confidence intervals show that the SRT$_N$ of list 10 is significantly different from the overall SRT$_N$, while the SRT$_N$ of list 11 is at the limit of a significant difference. In an attempt to avoid a situation with some test lists being singled out in clinical use as easier or more difficult than the others, list 10 and 11 were omitted from the test. The final 18 lists can be found in the appendix B. The appendix also contains seven practice lists compiled from sentences that were excluded during various stages of the development process. The sentences are equalized with respect to intelligibility, but the lists should only be used for practicing the test procedure.

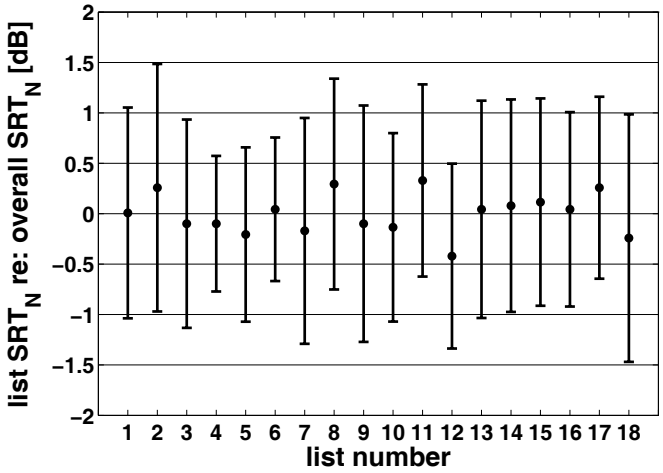


Figure 2.4: The list-SRT$_N$s relative to the overall mean (-3.15 dB). The bars indicate ± 1 standard deviation. The overall standard deviation is 1.0 dB; the standard deviation of the list-SRT$_N$s is 0.2 dB. The omission of list 10 and 11 from the original 20 lists means that list 10-18 in the present figure corresponds to list 12-20 in Fig. 2.3.

The following results are based on the final 18 lists. The mean SRT$_N$ across all lists and listeners is -3.15 dB with an overall standard deviation of 1.0 dB. A 2-way ANOVA shows a significant variation between listeners [$F(13, 221) = 3.04$, $p = 0.0004$], but no significant variation between lists [$F(17, 221) = 0.64$, $p = 0.86$]. The mean SRT$_N$ for each list relative to the overall SRT$_N$ is shown in Fig. 2.4. Here, lists 1-9 are the same as in the original set (Fig. 2.3), while lists 10-18 correspond to lists 12-20 in the original set. All list-SRT$_N$s lie within ± 0.5 dB of exact equality.

2.3.3 Test reliability

The reliability of a single SRT$_N$ determination with one test list can be estimated from the repeated SRT$_N$ measurements for the 14 listeners in the verification test. For each listener, an SRT$_N$ was calculated as the mean across the 18 lists. This SRT$_N$ value was then subtracted, listener by listener, from the individual SRT$_N$ determinations with the 18 lists. These differences can be regarded as the deviation of each single SRT$_N$

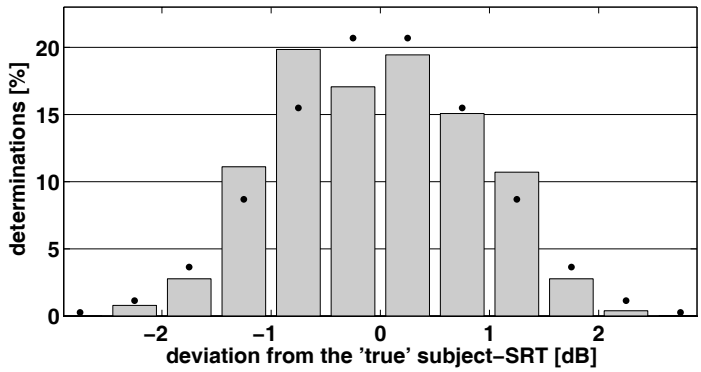


Figure 2.5: Percentage of SRT_N determinations with a given deviation from the individual SRT_N means. Bars show the percentage of measurements with the indicated deviation. Dots show the expected distribution under the assumption that results are normally distributed (std = 0.92 dB). Bin size is 0.5 dB.

assessment from the "true" SRT_N for the listener. Fig. 2.5 shows these deviations (n = $18 \cdot 14 = 252$) collected in bins of 0.5 dB. 71% of the deviations are within ± 1 dB of the "true" SRT_N. 93% are within ± 1.5 dB and 99% are within ± 2 dB. The within-subject standard deviation in the verification test was 0.92 dB; the dots in the figure indicate a normal distribution curve with this standard deviation. The empirical distribution shows a tendency for more deviations in the interval -1.5 dB to 1.5 dB than predicted by the theoretical distribution, but fewer measurements with deviations above ± 1.5 dB.

2.3.4 Phone distribution

The 20 original lists in the present study were created with respect to phonetic balance. After exclusion of two lists, the overall phonetic distribution was recalculated for the final 180 sentences and the deviation from this optimal distribution was determined for each of the final 18 lists. The sentences consist of 3363 phones in total; 1291 vowels (38%) and 2072 consonants (62%). The distribution of phones is listed in Table 2.1. 50 target values were defined for each list: 28 vowel counts, 20 consonant counts, one "stød" count, and one syllabic consonant count. The distribution of the deviations

Vowel distribution				Consonant distribution			
i	3.6%	œ	0.1%	p	1.2%	ʃ	0.1%
y	0.5%	ɑ	3.5%	t	2.5%	ʁ	2.1%
u	1.5%	ɒ	0.2%	k	1.5%	ʁ	1.4%
e	3.8%	iː	0.9%	b	3.0%	h	2.9%
ø	0.4%	yː	0.3%	d	5.7%	l	5.6%
o	0.7%	uː	0.4%	g	4.4%	j	1.3%
ə	3.0%	eː	0.4%	f	1.8%	w	0.7%
ɛ	2.3%	øː	0.2%	v	2.6%	m	4.0%
œ	0.3%	oː	0.8%	ð	4.2%	n	8.4%
ʌ	2.3%	ɛː	0.1%	s	6.9%	ŋ	1.3%
ɔ	1.1%	ɔː	0.9%				
æ	0.9%	æː	1.8%				
ɐ	3.3%	ɑː	0.7%				
a	4.0%	ɒː	0.5%				

Table 2.1: Average distribution of phones in the final sentence lists (180 sentences). The notation is in accordance with the International Phonetic Alphabet (IPA). During the trial-and-error process the goal was to reach a similar distribution for each list of 10 sentences. In addition to the listed phones, the number of "stød" and number of syllabic consonants were also included as counts in the optimization.

between the target values and the actual values of the 900 counts ($50 \cdot 18$) is shown in Fig. 2.6. 81% of the deviations are within ± 1, which can be compared to 58% in Nilsson et al. (1994), 75% in Vaillancourt et al. (2005), and 70% in Hällgren et al. (2006).

2.3.5 Psychometric function

The psychometric function of the test, shown in Fig. 2.7, was determined based on the responses across all listeners and lists during the verification test of the original 20 lists. During the test, sentences were presented at levels from -8 to 2 dB SNR in steps of 2 dB. At each level, the percentage of words/sentences that were correctly repeated was calculated. The first sentence in each list was not included in the statistics, as this sentence was presented several times to each listener. The steepest slope of the fitted cumulative normal distribution curve is 18.7 %/dB for sentences and 14.9 %/dB for the word-based curve.

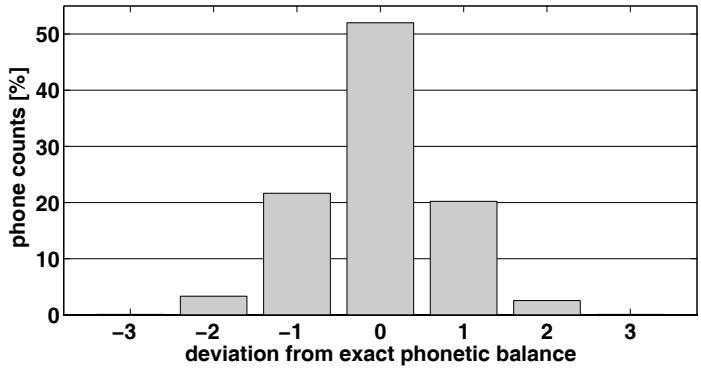


Figure 2.6: The distribution of the deviation between the target value and the actual value of the phonetic counts. As the target values are (normally) non-integers, the deviations are pooled in bins of size ± 0.5 around the indicated deviation.

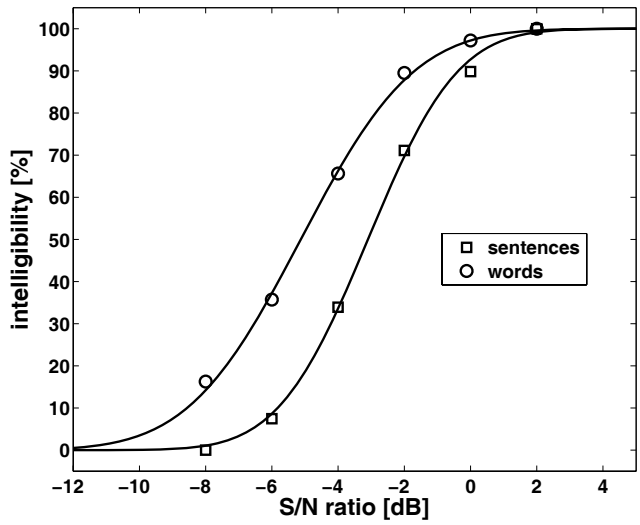


Figure 2.7: The psychometric function of the test based on correctly repeated sentences (squares) and correctly repeated words (circles). The solid curves are best fit normal distribution functions.

The slope of the psychometric function is to some extend influenced by the differences in the SRT_N between listeners. To investigate the influence of these differences, a sentence-based psychometric function was fitted to each individual listener. The steepest slope of these functions varied from 15.5 %/dB to 26.5 %/dB. The average slope, 19.8 %/dB, was only slightly steeper than the slope for the overall psychometric function.

2.4 Discussion

2.4.1 Comparison with other sentence tests

Talker, pronunciation, word frequency, language and sentence redundancy are some of the factors that influence the SRT_N of a speech intelligibility test. Nevertheless, the overall SRT_N of several HINTs falls within a narrow range: -2.9 dB for the American-English HINT (Nilsson et al., 1994), -3.3 dB for the Canadian-French HINT (Vaillancourt et al., 2005), -3.0 dB for the Swedish HINT (Hällgren et al., 2006), and -3.9 dB for the Cantonese HINT (Wong and Soli, 2005). The present test has an overall SRT_N of -3.15 dB. The standard deviation of the SRT_N across all lists and listeners is also similar for the different tests. The standard deviation amounts to 1.0 dB for the present test, 1.1 dB for the Canadian-French (Vaillancourt et al., 2005), 1.1 dB for the Swedish (Hällgren et al., 2006), and 1.7 dB for the Cantonese version (Wong and Soli, 2005). These standard deviations are dominated by the standard deviations within listeners, but are also influenced by the variation between the list-SRT_Ns (means across listeners). This variation was depicted in Fig. 2.4, and the figure can be compared to the corresponding figures in previous studies, e.g., Nilsson et al. (1994); Vaillancourt et al. (2005); Hällgren et al. (2006). The list-SRT_Ns in these studies are distributed within the interval ± 1 dB, while in the present study, the SRT_Ns are confined to the interval ± 0.5 dB. This low variability is partly caused by the exclusion of the two most diverging lists and the inclusion of eight sentence levels in the SRT_N calculation. It does, however, also imply that the equalization procedure of the present study has led to more homogeneous sentence intelligibilities than the HINT equalization procedure. This is presumably due to the fact that the

HINT procedure is based on an equalization of the average word intelligibility, which does not necessarily ensure equalized sentence intelligibilities.

The conflict that arises from using word scoring in the equalization procedure but sentence scoring in the adaptive test procedure cannot be resolved by switching to word scoring in the latter. A reliable and sensitive test with word scoring would require an equalization of the individual word intelligibilities - not only the average word intelligibility for each sentence - and this is in general not achievable for natural sentences.

2.4.2 Prediction of SRT_N improvements

The empirical results of the verification showed that 99% of the SRT_N determinations are within ± 2 dB of the listener's "true" SRT_N. In practice, in a laboratory or a clinic, the SRT_N will often be measured twice for a listener, once for each of two conditions. The purpose can be to compare two different hearing aids and to decide whether there is a significant difference. The null hypothesis of such a setup is that the SRT_N is equal in the two conditions, and the rejection of the hypothesis depends on the difference between the two SRT_N determinations. If the within-subject standard deviation of 0.92 dB is also assumed to be valid for hearing-impaired listeners, the deviation of the difference is $\sqrt{2} \cdot 0.92$ dB = 1.30 dB. In the case of a two-tailed test (there is no expectation about which condition will result in the lowest SRT_N), the 5% critical region will be limited by $\pm 1.96 \cdot 1.30$ dB = ± 2.5 dB. In the case of a one-tailed test (there is an expectation about which condition will have the lowest SRT_N), the 5% critical region will be limited by $1.64 \cdot 1.30$ dB = 2.1 dB. In practical use, this means that an SRT_N difference of 2.5 dB indicates a significant difference between two conditions for both one-tailed and two-tailed tests. The test must be performed with the same listener in the two conditions.

2.4.3 The HINT versus the CLUE equalization procedure

In the HINT equalization procedure, a linear relationship between SNR and intelligibility is assumed. A common assumption is that a 10% deviation in intelligibility is compensated by a 1 dB change of the SNR (an overview can be found in Soli and

Wong, 2008). This method of adjusting the intelligibility is rather coarse and does not take the changing slope of the psychometric function into account. At the 50% point of the function, the intelligibility will be more sensitive to changes in the SNR than when the intelligibility approaches 0% or 100%. The HINT method also makes the implicit assumption that a change in the SNR leads to the same change in intelligibility for all sentences, although this is probably incorrect. In contrast, the CLUE equalization procedure presented in the present study leads directly to an adjustment of the sentence RMS level, without any assumed relationship between intelligibility and SNR.

The sentence equalization process is a time consuming part of the development of a speech intelligibility test. The time efficiency of the equalization procedure can be estimated by how many times each sentence has been presented to a person during the process. During the HINT procedure, each sentence presentation involves a listener and an instructor, who need to go through the sentence two times: the initial presentation and the repetition by the listener. This equals to four "person-presentations". In Vaillancourt et al. (2005), each sentence was tested 36 times, resulting in 144 person-presentations per sentence. In Hällgren et al. (2006), the sentences were tested 32 times, resulting in 128 person-presentations per sentence. During the equalization of the present study, the sentences were, on average, presented 3.1 times to 18 listeners without an instructor present, resulting in 56 person-presentations per sentence. This means that the time consumption of the equalization process in the present study was more than halved compared to the procedure of previous HINT studies.

2.4.4 Limitations of the CLUE equalization procedure

One potential problem of the CLUE equalization procedure is that listeners will sometimes press "ok" for sentences that they have misunderstood or not heard fully. This is possible because there is no instructor to evaluate what the listener has perceived. When this happens repeatedly for the same sentences, it is an indication of an intelligibility flaw. During a pilot test before the final verification, some of the sentences were never correctly repeated. This was regarded as unacceptable and a screening was carried out in order to identify such sentences. These sentences were omitted from the final test lists. The sentences should have been identified earlier in the process and definitely before the equalization. A suggested procedure would be

to include a sentence screening at an early stage. All sentences that do not have an intelligibility of 100% at a relatively high SNR (e.g., 0 dB) should then be discarded.

During the equalization process, the sentences were presented in sequences that consisted of fewer and fewer sentences until all had been judged “ok”. The last sequences inevitably consisted of only a few sentences that became quite well-known to the listeners, who sometimes got locked in an “easy” judgement of these. Listeners were instructed to avoid pressing “easy” repeatedly for these sentences, but in practice this was unavoidable to some extent and the level was often lowered substantially. As this “false” lowering of the level would change randomly among the sentences from listener to listener, the effect has had only a minor influence on the average level adjustment. The problem could have been avoided by imposing a limit of four presentations per sentence. At the fourth presentation, the listener has either adjusted the sentence to a level outside of the acceptable ± 3 dB band or reverted to a level that previously was judged “easy” or “difficult”.

2.4.5 Characteristics of the speech material

The speech material in the present test has some characteristics that deviate from comparable sentence tests. The talker has a somewhat less clear pronunciation than a professional talker. This was to some extent the intention of choosing a non-professional talker: The test can thereby reveal the problems that some listeners have in a conversation, because they cannot interpret speech that is not clearly pronounced. The robustness of the test is, however, also affected, as the effect of the pronunciation will tend to occur in an uncontrolled manner at unspecified locations in the stimuli. The reasoning behind the decision to use a non-professional talker may therefore be doubtful. Also, the talker does not keep a constant quality of voice in all sentences. This is an unintended effect caused by a split of the recordings in several takes. The influence of this effect is, however, expected to be minor.

The sentence material contains sentences that can be perceived as more complicated and “unnatural” compared to other sentence tests. This is based on the fact that the sentences originated from a written source and that they did not undergo an evaluation for naturalness by a group of native speakers. In spite of this, the verification of the test with normal-hearing listeners showed a within-subject deviation

between SRT_N assessments that is equal to or lower than the deviation in comparable tests.

2.5 Conclusion

A speech intelligibility test with 18 sentence lists has been produced. The test is in many aspects comparable to HINTs developed for other languages and it has a similar overall SRT_N (-3.15 dB) and standard deviation (1.0 dB). The deviations of the list-SRT_Ns from the overall mean are less than 0.5 dB, and thus considerably lower than in the HINTs. The CLUE equalization procedure might be the reason for this. In future developments of sentence-based speech tests, it is therefore suggested to consider this method instead of using the traditional HINT equalization procedure.

Acknowledgements

We wish to thank the following colleagues for assistance and advice during the project: Preben Dømler, Nina Grønnum, Birgit Hutters, and Niels Reinholt Petersen (INSS, University of Copenhagen); Claus Lynge Christensen (ODEON); Carl Ludvigsen and Erik Schmidt (Widex); Thomas U. Christiansen, Torben Poulsen, Jørgen Rasmussen, and Eric R. Thompson (Acoustic Technology, DTU). We would also like to thank two anonymous reviewers for their helpful and constructive comments and suggestions. Finally, we would like to thank Oticon's research unit, Eriksholm, for feedback on the final test. The present work was supported by the Oticon Foundation and the H.C. Ørsted Foundation.

Chapter 3

The Danish Hearing in Noise Test (HINT)

This chapter is based on Nielsen and Dau (2009a)

Abstract

A Danish version of the Hearing in Noise Test (HINT) was developed. The test consists of 10 test lists and 3 practice lists, each including 20 sentences. The speech material is based on the Conversational Language Understanding Evaluation (CLUE) test (Nielsen and Dau, 2009b), which contains sentences that were equalized with respect to sentence intelligibility. In the present study, the sentences were evaluated for naturalness by a panel of 10 native Danish speakers and the sentences with the highest score were selected for the test lists. The practice lists were compiled from the remaining sentences. The test lists were validated with 16 normal-hearing and 16 hearing-impaired listeners. The normative speech recognition threshold in noise (SRT_N) obtained for NH listeners was -2.52 dB with a standard deviation of 0.87 dB. The within-subject standard deviation was 0.86 dB for NH listeners and 0.92 dB for HI listeners. This indicates an almost equal test reliability for these two groups. Retests after three weeks with both NH and HI listeners indicate that reliable results can be obtained also when sentence lists are reused with the same listener.

3.1 Introduction

Speech intelligibility tests using natural speech material are now available in several languages. A Dutch test (Plomp and Mimpen, 1979) was the first to assess the speech recognition threshold (SRT) by an adaptive procedure using everyday sentences with equalized intelligibility. The concept was further developed by Nilsson et al. (1994), who created the American-English Hearing in Noise Test (HINT). The HINT has since been developed in many other languages, e.g., Canadian-French (Vaillancourt et al., 2005), Cantonese (Wong and Soli, 2005), and Swedish (Hällgren et al., 2006). In addition, some researchers have developed tests that are very similar to the HINT but deviate in some minor aspects; examples are the French Intelligibility Speech Test (FIST; Luts et al., 2008) and, in Danish, the Conversational Language Understanding Evaluation (CLUE; Nielsen and Dau, 2009b). Both tests deviate from the HINT by the procedure employed for equalizing the intelligibility of the test sentences. The HINT represents one of the most common speech intelligibility tests, partly because of its availability on a commercial computer system that can administer the test. The system permits the inclusion of sentence and speech materials in various languages as long as these have been developed according to some basic criteria. The current HINT test procedure, as implemented in the system, has changed since the original HINT (Nilsson et al., 1994) was developed. A major change is the use of 20-sentence lists (instead of 10-sentence lists) and a corresponding increase in the number of sentences that are included in the calculation of the SRT. The original 10-sentence lists created in many language versions of HINT have been paired to 20-sentence lists in order to comply with the current test procedure.

The typical speech intelligibility assessment using the HINT is done in the presence of a background noise and the result is the *speech recognition threshold in noise* (SRT_N), which is the signal-to-noise ratio (SNR) where a predefined speech intelligibility of 50 % is achieved. When measured with a group of normal-hearing (NH) listeners, the average SRT_N, the *normative* SRT_N, lies between -5.3 to -2.6 dB for the majority of HINTs (see review of Soli and Wong, 2008). While a low SRT_N may indicate that the sentences of the test are clearly pronounced or have a simple and natural vocabulary, it is not the goal to achieve as low a normative SRT_N as

possible when developing a HINT. Instead, the goal is rather to create a test with a normative SRT_N that resembles the signal-to-noise ratio (SNR) of difficult everyday-listening situations for NH listeners. For both NH and HI listeners, the presented speech stimuli in such a test will be balanced around SNRs that the listener is able to manage in everyday life. The SRT_N is thus expected to be a good predictor of the listener's abilities in situations where speech communication is difficult. A "realistic" SRT_N is also important for assessing the efficiency of hearing aids, particularly when investigating more advanced features such as noise reduction algorithms. At very low SNRs, these algorithms are typically not able to improve speech intelligibility; therefore, they cannot be tested using speech tests that produce very low SRT_Ns. The HINT effectively raises the SRT_N by using sentences that are unique and unpredictable and by employing sentence scoring (instead of word scoring).

Nielsen and Dau (2009b) developed a speech intelligibility test that is similar to the HINT. The test deviated in some methodological aspects, therefore it was named *Conversational Language Understanding Evaluation* (CLUE). The main deviation is the procedure used for equalizing the intelligibility of the test sentences. The equalization of the sentence intelligibilities is crucial in a sentence test, otherwise the test will produce unreliable SRT_N assessments and be insensitive to small differences in these thresholds (MacLeod and Summerfield, 1990). The standard HINT development process includes an equalization procedure that leads to an equalization of the average *word* intelligibility of each sentence (e.g., Nilsson et al., 1994; Soli and Wong, 2008). However, an equalization of the average word intelligibility for each sentence does not necessarily ensure equalized *sentence* intelligibilities (Nielsen and Dau, 2009b), although this is the objective of the procedure. The equalization procedure in Nielsen and Dau (2009b) is instead based on subjective intelligibility assessments of each sentence as a whole by two groups of NH listeners. The procedure has the additional advantage that it does not require an assumption about how much the RMS level of a sentence needs to be adjusted in order to compensate for an intelligibility deviation (the performance-intensity function). The CLUE test consists of 18 phonetically balanced 10-sentence lists (Nielsen and Dau, 2009b). The overall SRT_N with 14 NH listeners is -3.15 dB with a standard deviation of 1.0 dB. The mean SRT_N of all test lists is within ± 0.5 dB of the overall SRT_N and there are no significant differences

between the lists. These results are comparable to or better than those of HINTs in other languages, i.e., it seems that the equalization procedure used in CLUE has equalized the sentence intelligibilities more effectively than the procedure used in HINT.

The CLUE test was presented to the Danish hearing aid companies Oticon, GN Resound, and Widex. One of the companies conducted an evaluation of the test in order to decide whether to include it in the company's test battery. A major interest of the evaluation was also how homogeneous the test material would be with HI listeners. A problem with inhomogeneity for HI listeners - although of equal difficulty when validated with NH listeners - has been reported for the QuickSIN lists (McArdle and Wilson, 2006).

The evaluation (personal communication) acknowledged that the validation results for CLUE were comparable to those of HINTs in other languages. However, several potential concerns were raised regarding (i) the speech material, (ii) the choice of the talker, and (iii) the scoring rules, as outlined in the following.

(i) The CLUE sentences were based on written materials like newspapers and magazines and fulfilled a set of criteria (Nielsen and Dau, 2009b). In contrast, typical HINT sentences have been taken from oral materials aimed at children. The sentences in CLUE were not formally evaluated in terms of naturalness by a panel of native Danish speakers. As a result, the sentences may vary in naturalness and level of abstraction. The sentences in CLUE tend to be more complicated than the typical HINT sentence, e.g., "De talte lidt om fremtiden" (they talked a bit about the future). Some sentences have inversion (reversed word order), e.g., "Under bogen ligger en tegning" (under the book lies a drawing) and some verbs are in passive form, e.g., "Cykler kan lejes mange steder" (bikes can be rented in many places). In the evaluation, some words and expressions were considered "old-fashioned", e.g., "Skoledrengen drikker et glas mælk" (the schoolboy drinks a glass of milk).

(ii) The talker of the CLUE sentences was a 38 year old male with a background in phonetics, but with no previous experience in speech recordings. The evaluation pointed out that the talker's voice quality varies and that his pronunciation is "remarkable" partly because of tension. The pronunciation of some sentences is less clear and

the speed of speech varies. It was argued that a trained talker would be preferable in order to have a speech material as consistent and “transparent" as possible.

(iii) The scoring rules for a sentence test, which determine when a listener’s response to a sentence is considered to be correct, normally permit some variations in the accepted response. The CLUE scoring rules permit both some general variations and a few specific variations. The general variations, such as change of verb tense, are similar to those of the HINTs in the other Scandinavian languages, Norwegian and Swedish. The specific variations, such as “vi/de" (we/they) are unique to CLUE. (The Norwegian HINT permits other specific variations.) It was argued in the evaluation that the CLUE scoring rules might cause less consistent scoring by the test leader than necessary and a clarification was recommended. It was, however, also acknowledged that in a sentence test, the scoring of the responses will always, to some extend, be influenced by the judgements of the test leader.

A collaboration between the three Danish hearing aid companies and the Centre for Applied Hearing Research (CAHR) was established with the objective of creating a new speech intelligibility test that was based on CLUE, but took the concerns mentioned above into account. The objective was to validate the test for both NH and HI listeners. Furthermore, the goal was to create a test that corresponds to the current HINT standard (Bio-logic Systems Corp., 2005), so it can be referenced as “the Danish HINT". The standard requires that each of the new test lists contains 20 sentences. It was assumed that a satisfactory speech material could be achieved by exchanging some of the CLUE test sentences with sentences from the CLUE practice lists.

3.2 From CLUE to Danish HINT

3.2.1 Test of naturalness

The naturalness of the sentences was judged by a panel of 10 native and “naive" Danish speakers and by two professional linguists. For various reasons, 15 of the CLUE practice sentences were rejected in advance, leaving 235 CLUE sentences for the naturalness test. The panel judged the written version of the sentences on a

scale from 1 (= “artificial") to 7 (= “natural"). The requirements for a sentence to be “natural" were (translated from Danish) that it does not contain unusual Danish words, and that it could have been used in an ordinary conversation. Sentences with a mean rating of 5 or above among the naive participants were accepted for the test. In addition, up to three sentences per test list with a score between 4.0 to 4.9 were permitted. These criteria were less strict than in other HINT tests, where the criterion has been a score of 6 for a sentence to be accepted. A score of 5 or above was achieved by 176 sentences, and 41 sentences received a score between 4 and 5. A sufficient number of “natural" sentences was thus available to compile 10 new 20-sentence lists.

3.2.2 Generation of the test lists

The 18 original CLUE test lists and two of the CLUE practice lists were combined to create 10 20-sentence lists. The CLUE list with the lowest SRT_N was successively paired with the list with the highest SRT_N in an attempt to achieve lists with equalized SRT_Ns. The process of exchanging the “unnatural" sentences present in these lists with sentences from the pool of “natural" sentences was conducted by a computer-based trial-and-error routine in order to obtain a phonetic distribution for each list as close as possible to the overall distribution for the 200 sentences. As in Nielsen and Dau (2009b), the distribution was based on 28 vowels, 20 consonants, the Danish glottal stop “stød", and syllabic consonant, representing in total 500 distribution counts (50 parameters times 10 lists). In the final lists, 88 % of the deviations between the target values and the actual values of the counts were within ± 2; the largest deviation was 4, which only occurred for one of the 500 counts. The 24 sentences with a naturalness score of 4.0 to 4.9 were distributed evenly among the 10 sentence lists.

The 10 test lists are listed in Appendix C. Included are also three practice lists that were compiled from the sentences that were deemed “unnatural" or omitted at previous stages.

3.2.3 Allowed response variations

A new set of rules for permitted variations in the listener response were created for the Danish HINT. The only major difference to the CLUE scoring rules was the omission of the specific word substitutions permitted in CLUE. The Danish HINT rules also contain a few clarifications. The following response variations are accepted: 1) Change in verb tense; 2) change in article; 3) change between singular and plural nouns; 4) reordering of words; 5) addition of extra words or phones; 6) omission of a single phone (e.g., the [t] that changes adjectives to adverbs in Danish). Several variations are permitted in a single response.

3.3 Test validation with NH listeners

The purpose of the validation with NH listeners was to establish normative data for the test and to investigate whether the test performance is equivalent to that of HINT in other language versions (an overview of performance data can be found in Soli and Wong (2008)). The performance measures included the mean SRT_N and standard deviation across listeners and lists, the mean SRT_N for each list with standard deviation across subjects (list equivalence), and the sensitivity, i.e., slope of the psychometric function. Furthermore, the test-retest variance was measured.

3.3.1 Method

Listeners

Sixteen (8 male, 8 female) NH listeners participated in the investigation. Their age was between 19 and 43, mean 33.6 years. The requirements for participation were: 1) Age 18-45 years; 2) hearing threshold level $\leq$ 20 dB HL at both ears; 3) hearing threshold level of 25 dB HL allowed at one frequency per ear; 4) Danish as native language; 5) no previous experience with CLUE; and 6) variation in the educational background for the group.

Apparatus and procedure

The experiment took place in a sound-proofed booth and the stimuli were presented diotically over Sennheiser HD580 headphones. The use of headphones (as opposed to a sound-field setup) was in accordance with the typical HINT validation approach (Soli and Wong, 2008). In the present study, only the condition with speech and noise coming from the same direction (“front") was investigated. Audio files of all the listeners’ responses were recorded by a hard disk recorder. These recordings were kept for later investigation of how the test leader performed in following the scoring rules.

The validation tests were conducted according to the standard HINT procedure (Bio-logic Systems Corp., 2005). The noise level was fixed at 65 dB(A), whereas the speech level varied according to the adaptive test procedure of the HINT. The order of the sentences within each list was randomized before presentation of the list. The test procedure was implemented in a modified version of the MATLAB application used for the CLUE validation. The listener received oral instructions before the test. The listeners were encouraged to guess words or parts of sentences when they were in doubt during the test.

A practice round including two 20-sentence practice lists in noise was completed to familiarize the listener with the task and to make sure that the instructions were correctly understood. Each listener was then tested with all 10 test lists. The order of the test lists was counterbalanced across listeners (using Latin squares) to avoid order effects. A short break was included after completion of the first five lists.

A retest was conducted three weeks after the test. The retest followed the same schedule and procedure as used in the test. The order of lists for each listener was the same, but the randomization of the sentences within the lists was different.

3.3.2 Results

Validation and reliability

All SRT_Ns in the present study were calculated according to the current HINT standard (Soli and Wong, 2008). The overall SRT_N across all lists and listeners was -2.52 dB with a standard deviation of 0.87 dB. The within-subject standard

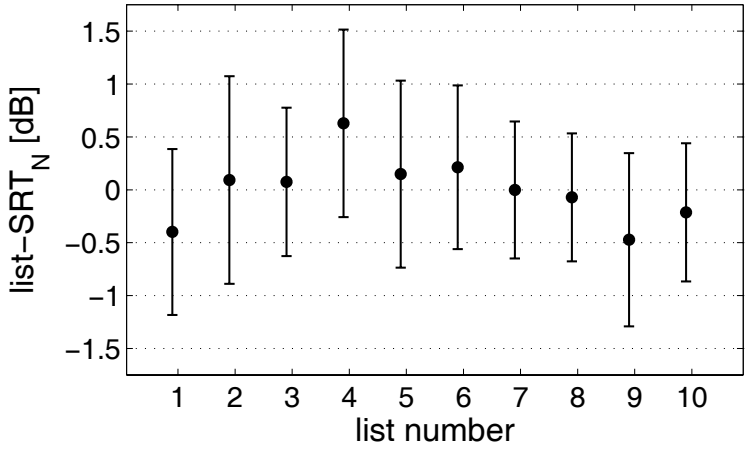


Figure 3.1: The list-SRT$_N$s relative to the overall mean (-2.52 dB) based on the validation test with 16 NH listeners. The bars indicate ± 1 standard deviation.

deviation was 0.86 dB. A two-way ANOVA showed a significant variation between lists [$F(9, 135) = 2.37$, $p = 0.016$], but no significant variation between listeners [$F(15, 135) = 1.34$, $p = 0.19$]. For each of the 10 lists, a list-SRT$_N$ was calculated as the mean across the 16 listeners. The result is shown in Fig. 3.1. The list-SRT$_N$ standard deviation is 0.32 dB and the maximum deviation from the overall mean is 0.63 dB. Figure 3.2 shows the mean SRT$_N$ across the 10 test lists for each of the 16 NH listeners. These subject-SRT$_N$s lie within an interval of 1.1 dB.

The reliability of the test can be judged by the variations that are observed for repeated measurements of the SRT$_N$ for the same listener. For each listener, the subject-SRT$_N$ was regarded as the "true" SRT$_N$ and the deviations between this value and the actual SRT$_N$ measurements were calculated for all listeners. 79 % of the measurements were within ± 1 dB of the "true" SRT$_N$; 94 % of the measurements were within ± 1.5 dB; and 99 % were within ± 2 dB. This is consistent with a normal distribution with the determined within-subject standard deviation of 0.86 dB.

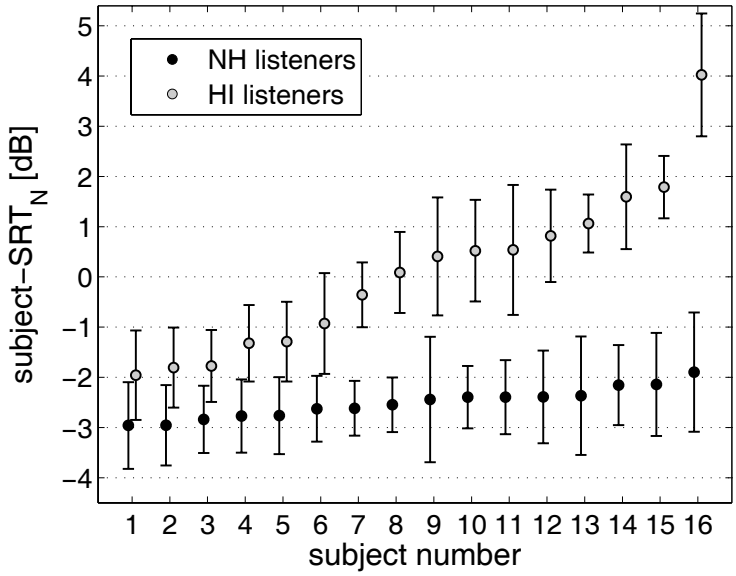


Figure 3.2: The absolute subject-SRT$_N$s measured during validation with 16 NH listeners (black circles) and validation with 16 HI listeners (grey circles; refer to section 3.4). For each group the listeners are sorted with respect to their mean SRT$_N$. The bars indicate ± 1 standard deviation.

Psychometric function

The psychometric functions of the test were determined for each individual listener. The data points were based on the percentage of correctly repeated sentences at each of the SNRs of the adaptive procedure. Sentences at list positions 5-20 in the 10 test lists were included in the calculation; these correspond to the sentences included in the calculation of the SRT$_N$. The SNR differences of 0.2 dB that are possible with the current HINT procedure were collapsed into integer values; all presentations in the interval [x-0.4; x+0.4] were considered to have been done at the same SNR. A cumulative normal distribution was fitted to each set of data points hereby estimating a psychometric function for each listener. The steepest slope of these curves varied from 10.9 to 20.7 %/dB; the mean value was 16.8 %/dB.

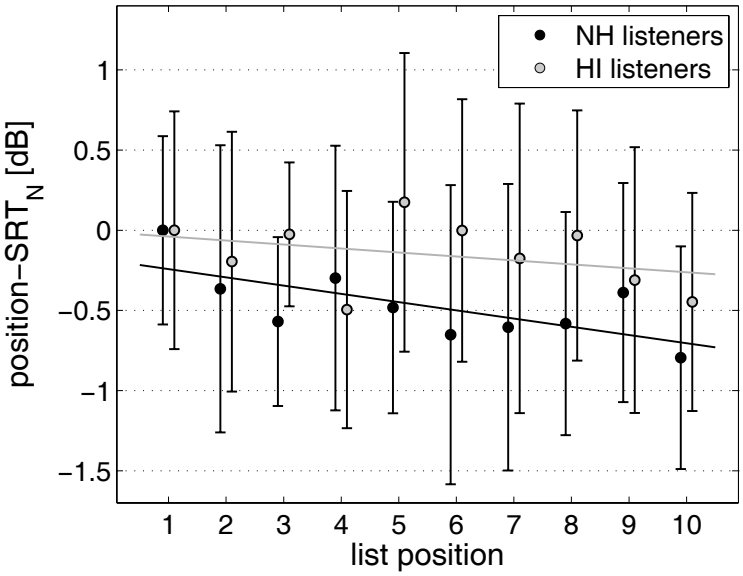


Figure 3.3: The mean SRT_N across lists and listeners as a function of the list position during the test session. Data are adjusted with respect to the SRT_N of position 1. Bars indicate one standard deviation. The black linear regression line is a best fit to the means for the NH listeners. The grey line is a similar fit for the HI listeners (refer to section 3.4).

Training effect

Figure 3.3 shows the mean SRT_N as a function of the list position during the test sessions. For each position, the SRT_N was calculated as the mean across the combinations of listeners and lists presented at that position (n = 16). The data were normalized with respect to list-SRT_N and subject-SRT_N, i.e., the effects of list and listener have been removed. A linear regression line was fitted to the data for all 10 list positions; the gradient was -0.05 dB/position corresponding to an SRT_N decrease of approximately 0.5 dB from the first to the last measurement. The major effect of the training seems to occur during the two first list presentations. When these two presentations were taken out of the linear regression, the gradient reduced to -0.027 dB/position.

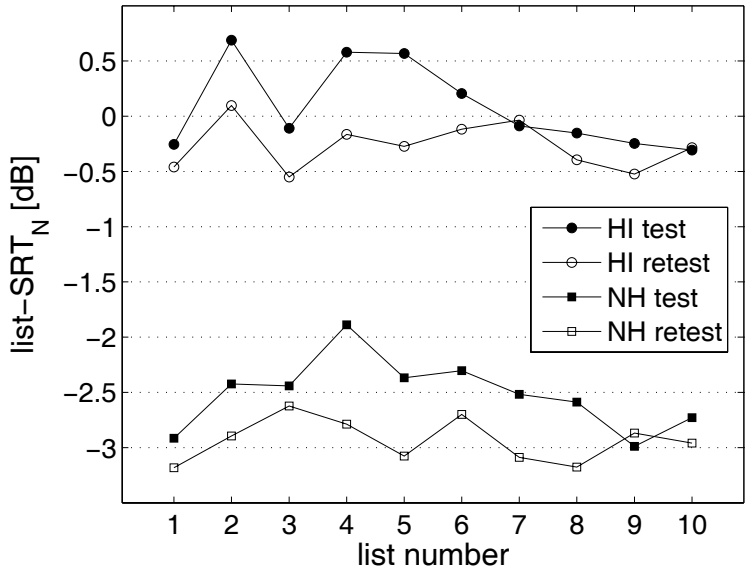


Figure 3.4: Comparison of the absolute list-SRT$_N$s in test and retest for the 16 NH listeners (solid and open squares) and for the 16 HI listeners (solid and open circles; refer to section 3.4).

Retest

The retest was conducted with all listeners after exactly three weeks. Except for three listeners, the visits took place at the same time of the day. The same data were gathered as during the preceding test and all calculations were done in a similar manner. All retest results are, however, based on absolute SRT$_N$s since any normalization with respect to list or listener would remove the retest effect that is under investigation.

The overall SRT$_N$ across all lists and listeners was -2.94 dB, a decrease of 0.42 dB from test to retest. The overall standard deviation was 0.75 dB and the within-subject standard deviation was 0.69 dB. A two-way ANOVA showed no significant variation between lists [$F(9, 135) = 1.31$, $p = 0.24$], but a significant variation between listeners [$F(15, 135) = 3.06$, $p = 0.0003$]. The list-SRT$_N$s were within $\pm$0.31 dB of exact equality.

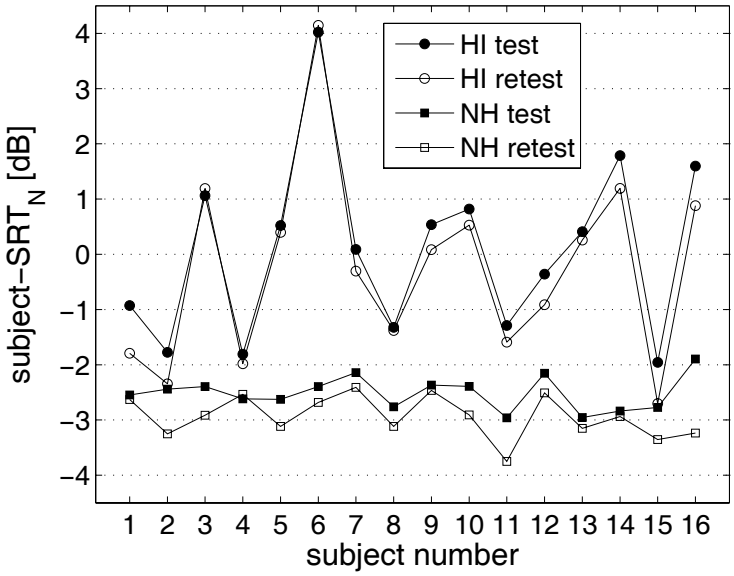


Figure 3.5: Comparison of the absolute subject-SRT$_N$s in test and retest for the 16 NH listeners (solid and open squares) and for the 16 HI listeners (solid and open circles; refer to section 3.4).

Figure 3.4 compares the average list-SRT$_N$ in test and retest; the mean decrease is 0.42 dB. The largest change occurs for list 4 with an SRT$_N$ decrease of 0.90 dB. Only for list 9 does an SRT$_N$ increase occur (0.12 dB).

Figure 3.5 is a comparison of the subject-SRT$_N$s. The SRT$_N$ changes between test and retest are relatively unevenly distributed. Listeners 2, 11 and 16 improve their performance by 0.8 to 1.3 dB, while listener 4 experiences an SRT$_N$ increase of 0.08 dB.

The SRT$_N$ as a function of the list position during test and retest is shown in Fig. 3.6. For each position, the SRT$_N$ was calculated as the mean across the combinations of listeners and lists presented at that position. The effect of training during the first test seems to last at least until the retest; the SRT$_N$ of the first retest list is close to that of the last test list. During the retest, the training effect was approximately half, -0.024 dB/position, of the effect during the first visit.

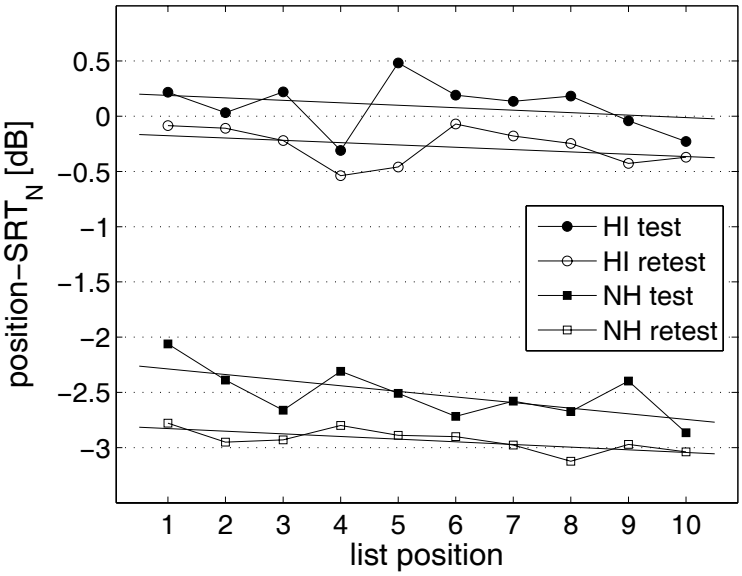


Figure 3.6: The mean SRT_N across lists and listeners as a function of the list position for the 16 NH listeners (solid and open squares). The straight lines are best fit linear regression lines to the means. Also shown are the results for the 16 HI listeners (solid and open circles; refer to section 3.4).

3.3.3 Discussion

The Danish HINT evaluated in this study has normative data that are comparable to the HINT in other language versions. The SRT_N for the 16 NH listeners is -2.52 dB, which falls slightly outside of range -5.3 to -2.6 dB observed for the 13 languages listed in Soli and Wong (2008). The average SRT_N for these languages is -3.9 dB. The relatively high SRT_N for the Danish test is presumably caused by the complexity of the sentences and the use of a non-professional talker. This does not necessarily represent a disadvantage of the test. A primary goal of creating a new test was to achieve an SRT_N that is considerably higher than that of existing Danish tests, e.g., the DANTALE II test (Wagener et al., 2003) with a normative SRT_N of -8.4 dB. The standard deviation of the SRT_N across lists and listeners was 0.87 dB, which is slightly below the mean of the standard deviations in Soli and Wong (2008).

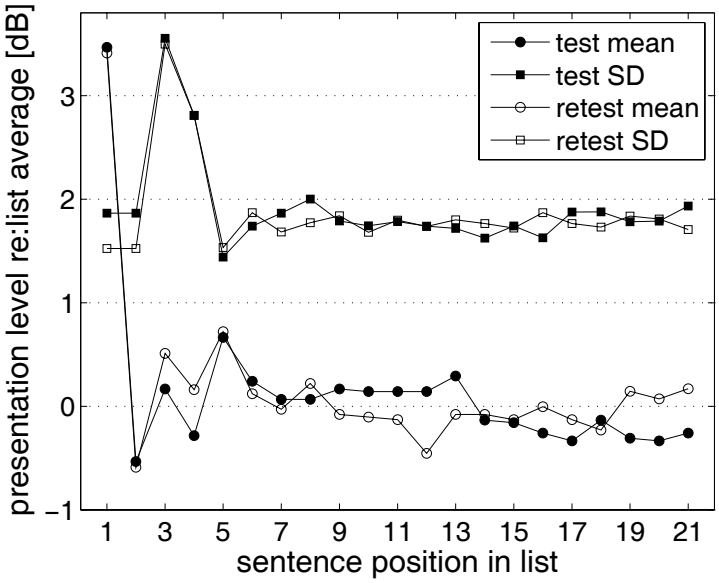


Figure 3.7: The presentation level mean and standard deviation across all lists and all NH listeners as a function of the sentence position in the list. The presentation level is given relative to the average level of sentence 5-21 in each list. Level 5 is the first level that is included in the SRT_N calculation.

A comparison of the test and retest results imply that the test is reusable after three weeks. The statistical data for the retest are better than for the initial test. The within-subject standard deviation is reduced and the significant difference between the lists that was observed in the test is not observed in the retest. The decrease of the overall SRT_N of 0.42 dB is too small to affect the functionality of the test. It is recommended, however, to keep track of what status (not heard before or heard before) the lists have for each listener, and as far as possible use lists with the same status during the same investigation.

The normative SRT_N of the present test (-2.52 dB) is 0.6 dB higher than for the CLUE test (-3.15 dB). An analysis of the underlying data shows that the difference is highly significant. The changed adaptive procedure of the HINT compared to CLUE may explain part of this increase. This is clear from Fig. 3.7 where the mean

presentation level across all lists and listeners is calculated relative to the average level of sentence 5-21 in each list. (Sentence 21 does not exist but its level follows from the response to sentence 20.) The HINT procedure tends to produce an “overshoot" of approximately 0.7 dB for the first sentence level that is included in the SRT_N calculation (sentence position 5). The overshoot is caused by the high presentation level for sentence 1. In CLUE, there was a tendency for the first level of the SRT_N calculation to show a slight undershoot (Nielsen and Dau, 2009b). This difference is one factor that raises the normative SRT_N for the Danish HINT compared to CLUE.

The difference in professional background and age of the NH listeners in the CLUE validation experiments and the present experiments may have had an influence on the normative SRT_N. The NH listeners in the CLUE validation were mainly students from the Technical University of Denmark, with a mean age of 22.9 years. In the present study, the listeners had varying professional backgrounds and a mean age of 33.6 years. The age difference of more than 10 years may represent a (slightly) reduced ability to detect speech in noise. Some of the HINT listeners also felt less relaxed in the test situation than the students did in the CLUE validation. This may to some extend have affected their concentration.

The change in the scoring rules between CLUE and HINT might also explain part of the normative SRT_N increase. When fewer response alternatives are permitted, as in the HINT, the probability of an incorrect response increases and more upward steps are taken during the adaptive procedure. This leads to a higher SRT_N.

From CLUE to HINT, the overall standard deviation across lists and listeners was reduced from 1.0 dB to 0.87 dB. This reduction is smaller than expected if one HINT measurement (20 sentences) is considered equal to two CLUE measurements (2 times 10 sentences). Statistically, the mean of two CLUE measurements would only have a standard deviation of 1.0 dB/$\sqrt{2} = 0.7$ dB. The relatively high standard deviation for the HINT is related to the shallower psychometric function for the test compared to CLUE; the mean steepest slope for the psychometric functions of HINT was 16.8 %/dB and for CLUE 19.8 %/dB (Nielsen and Dau, 2009b).

3.4 Test validation with HI listeners

The Danish HINT was also validated with a group of HI listeners. Since a speech intelligibility test is primarily used for assessments involving HI listeners, a validation of the test with such listeners was considered important. The within-subject standard deviation of the SRT_Ns and the variation of the list-SRT_Ns were considered particularly important as they express the reliability of the test.

3.4.1 Method

Listeners

Sixteen HI listeners (10 male, 6 female) participated in this part of the validation. Their age was between 61 and 69, mean 65.9 years. The requirements for the listeners were: 1) Age 60-70 years; 2) hearing loss caused by presbyacusis, reflecting symmetrical mild to moderate sloping hearing loss; 3) at least one year experience with wearing a hearing aid; 4) Danish as native language; 5) experience with DANTALE II (the Danish Hagerman test); 6) no previous experience with CLUE; and 7) variation in the educational background for the group.

Apparatus and procedure

The test procedure for the HI listeners was similar to the procedure for the NH listeners. All testing was conducted without the use of hearing aids. The practice procedure before the actual test was extended from two to four lists and the two initial lists were presented in quiet. The practice in quiet was included in order to determine an appropriate noise level for the subsequent testing in noise. The speech recognition threshold in quiet (SRT_Q) of the second practice list determined the level of the noise in the subsequent test. If $SRT_Q \leq 45$ dB(A), the noise level was fixed at 65 dB(A). If $SRT_Q > 45$ dB(A), the noise level was fixed at SRT_Q + 20 dB(A). This determination of the level for HI listeners corresponds to the current HINT recommendations (Biologic Systems Corp., 2005). Two practice lists (the first practice list in quiet was reused) were presented in noise before continuing with the actual test in noise.

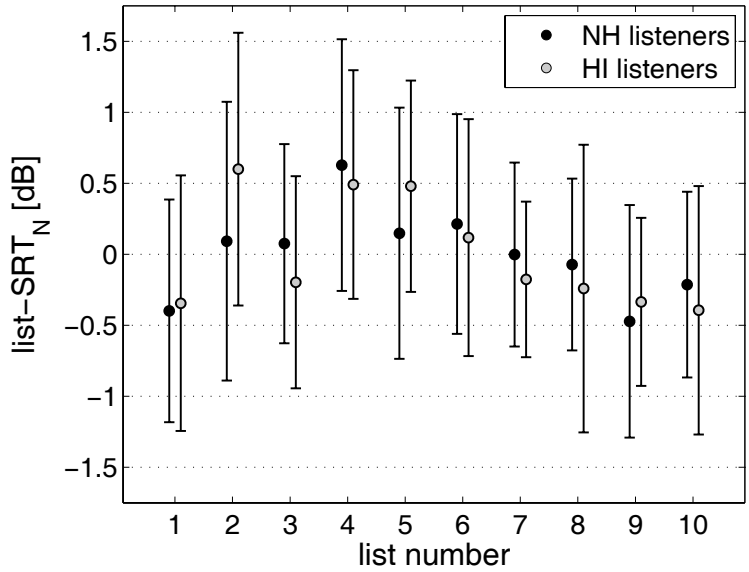


Figure 3.8: The mean list-SRT$_N$s across the 16 HI listeners (grey circles) compared to the mean list-SRT$_N$s across the 16 NH listeners (black circles). The bars indicate ± 1 standard deviation. The calculations are based on SRT$_N$s that are normalized with respect to the individual subject-SRT$_N$s.

A retest was completed three weeks after the initial test. The retest followed the same schedule and procedure as the test, except that the practice lists in quiet were not presented. The same individual noise levels as determined during the first visit were reused.

3.4.2 Results

Validation and reliability

The overall SRT$_N$ for the HI listeners was 0.09 dB with a standard deviation of 1.79 dB. The within-subject standard deviation was 0.92 dB. A two-way ANOVA showed a significant variation between lists [$F(9, 135) = 3.28$, $p = 0.0012$], and a very significant variation between listeners [$F(15, 135) = 35.31$, $p < 0.0001$].

As for the 16 NH listeners, the 10 list-SRT$_N$s were calculated across the 16 HI listeners; the result is shown in Fig. 3.8 in comparison to the NH listeners. For the HI listeners, the list-SRT$_N$ standard deviation is 0.39 dB and the maximum deviation from the overall mean is 0.60 dB. The list-SRT$_N$s are similar for the NH and the HI listeners; the largest deviation of 0.50 dB was observed for list 2. However, even for this list, an unpaired t-test did not show a significant difference between the list-SRT$_N$ for the HI and the NH condition [$p = 0.15$].

The mean subject-SRT$_N$s across the 10 test lists are shown in Fig. 3.2. The listeners are sorted according to their mean SRT$_N$. The variation between the subject-SRT$_N$s was much larger than for the NH listeners with the lowest value at -1.96 dB, and the highest value at 4.0 dB.

The test reliability is judged by the deviation between the individual subject-SRT$_N$s and the actual SRT$_N$ measurements for each HI listener (corresponding NH results in parenthesis): 74 (79) % of the SRT$_N$ measurements lay within $\pm$1 dB of the subject-SRT$_N$; 91 (94) % of the measurements within $\pm$1.5 dB; and 99 (99) % within $\pm$2 dB. These results reflect that the within-subject standard deviations for the HI and the NH listeners are similar.

Psychometric function

The individual psychometric functions for the HI listeners were determined as for the NH listeners. The slopes were based on a fitted cumulative normal distribution curve for each individual. The steepest slopes of the functions for all HI and NH listeners are shown in Fig. 3.9 as a function of the corresponding subject-SRT$_N$ for each listener. The steepest slopes for the HI listeners vary from 7.5 to 24.1 %/dB; the mean value is 14.7 %/dB. For the NH listeners, the slopes were 10.9 to 20.7 %/dB, with a mean value of 16.8 %/dB. For the HI listeners, there is a significant correlation of -0.65 between the slope of the psychometric function and the SRT$_N$ [$p = 0.006$]. For the NH listeners, there was no correlation [$r = -0.03$]. An unpaired t-test did not show a significant difference between the mean of the steepest slopes for the 16 NH listeners and the mean for the 16 HI listeners [$p = 0.15$].

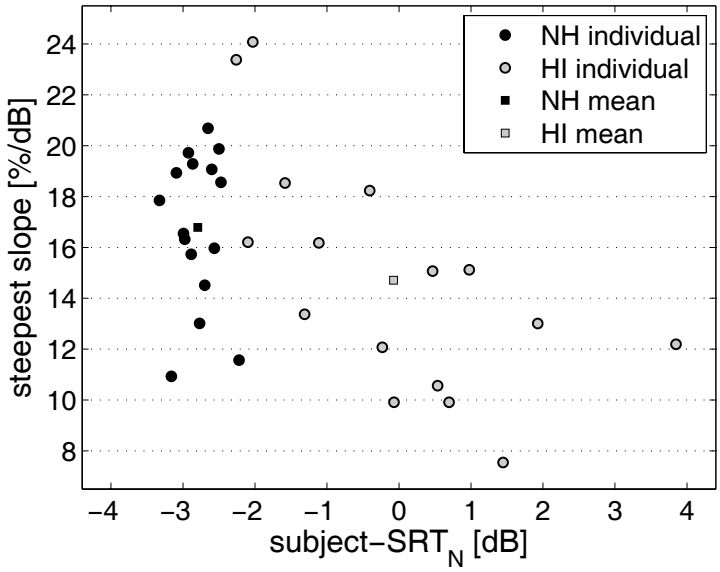


Figure 3.9: Steepest slope of the psychometric functions for the NH listeners (black circles) and the HI listeners (grey circles) as a function of the corresponding subject-SRT$_N$. The slopes are based on fitted cumulative normal distribution curves. The mean slope and mean subject-SRT$_N$ for the two groups are marked by squares.

Training effect

The SRT$_N$ as a function of the list position during the test session is shown in Fig. 3.3. For each position, the SRT$_N$ was calculated as the mean across all combinations of listeners and lists presented at that position (n = 16). The data were normalized with respect to the list-SRT$_N$s and the subject-SRT$_N$s, i.e., the effects of list and listener have been removed. A linear regression line was fitted; the gradient was -0.025 dB/position corresponding to an SRT$_N$ decrease of approximately 0.22 dB from first to last measurement due to training effects.

Retest

The retest was conducted for the HI listeners after three weeks. The overall SRT_N across all lists and listeners was -0.27 dB, a decrease of 0.36 dB from the initial test. The overall standard deviation was 1.86 dB and the within-subject standard deviation was 0.83 dB. A two-way ANOVA showed no significant variation between lists [$F(9, 135) = 1.09$, $p = 0.37$], but a very significant variation between listeners [$F(15, 135) = 44.6$, $p < 0.0001$]. The list-SRT_Ns were within ± 0.38 dB of exact equality.

Figure 3.4 compares the average list-SRT_N in test and retest; the mean decrease is 0.36 dB. The largest change occurs for list 5 with an SRT_N decrease of 0.84 dB. For two lists (7 and 10), a very slight SRT_N increase occurs.

Figure 3.5 is a comparison of the subject-SRT_Ns. Only for listeners 1, 2, 12, 14, 15, and 16 is there a change of more than 0.5 dB, whereas the changes are relatively small for the remaining listeners. The largest decrease is 0.87 dB (listener 1), which is less than observed for the NH listeners (1.3 dB, Fig. 3.5).

The absolute SRT_N as a function of the list position is shown in Fig. 3.6. During the retest there are level shifts in both directions and no obvious training effect. A fitted line shows a slight SRT_N decrease of 0.2 dB from the first to the last list presentation, similar to the effect during the initial test.

3.4.3 Discussion

The overall SRT_N for the HI listeners (0.09 dB) was 2.6 dB higher than for the NH listeners (-2.52 dB). This suggests that the test is sensitive to the listeners' ability to follow a conversation in noise. The overall standard deviation was also larger for the HI listeners (1.8 dB) compared to the NH listeners (0.87 dB). This increase is primarily caused by the between-subject variation of the SRT_N and indicates that the test is able to differentiate listeners with respect to hearing capabilities. The SRT_Ns are evenly distributed over the interval -2 to 2 dB, with one listener having an exceptional value of 4 dB (Fig. 3.2).

The stability and reliability of the test can be judged from the within-subject standard deviation. The deviation, 0.92 dB, for the HI listeners is only marginally

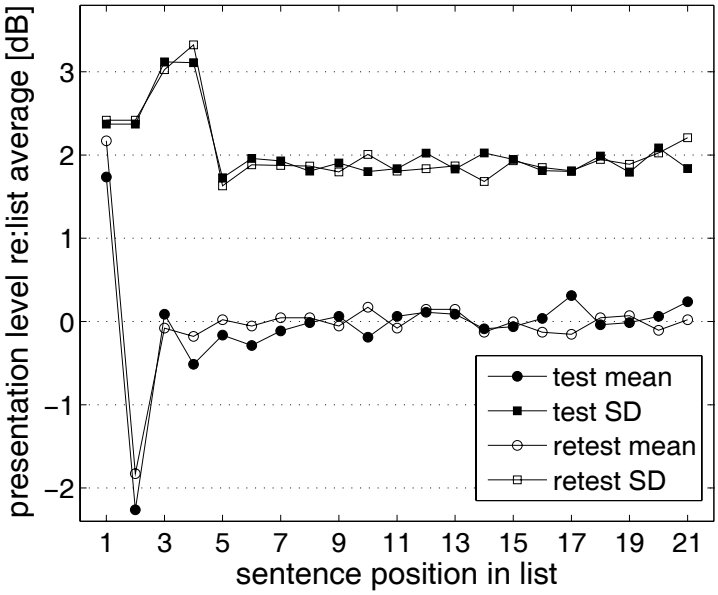


Figure 3.10: The presentation level mean and standard deviation across all lists and all HI listeners as a function of the sentence position in the list. The presentation level is given relative to the average of sentence level 5-21 for each list. The level of sentence 5 is close to the average presentation level for the remaining sentences.

larger than the value of 0.86 dB for the NH listeners. The reliability of the test is thus almost the same for HI listeners as for NH listeners. However, the reliability of the test for the HI listeners may be this high partly because of their experience from participation in listening tests with DANTALE II; such an experience was a requirement for their participation in the present study. Trained listeners are typically more focussed on the task and show a more reliable performance, which effectively reduces the within-subject standard deviation. The experience of the HI listeners may also explain why the training effect during the test was smaller for this group than for the NH listeners, and why the difference of the results between test and retest is smaller for the HI than for the NH listeners.

The evolvement of the presentation levels through the sentence lists is shown in Fig. 3.10. The level is stable from position 5 and onwards. The “overshoot" effect that was seen for level 5 for the NH listeners does not occur here. The adaptive procedure of the HINT thus seems more appropriate for testing with HI listeners than with NH listeners.

The retest showed a decrease in the overall SRT_N of 0.36 dB compared to the initial test. This is slightly less than the decrease for the NH listeners (0.42 dB). The decrease is relatively evenly distributed; none of the list-SRT_Ns or subject-SRT_Ns decrease with more than 0.9 dB (Fig. 3.4 and Fig. 3.5). This suggests that the test can be used after a break of three weeks. Still, as for the NH listeners, it is recommended only to use lists with the same status within one investigation. The reliability of the retest results was actually better than that of the initial test results. This is reflected in a slightly reduced within-subject standard deviation and reduced between-list variation. This phenomenon was also observed for the test/retest of the NH listeners and is probably explained by training effects.

3.5 Effects of learning and memory

A third experiment was performed with an additional group of HI listeners. The purpose was to investigate the training effect separated into learning and memory. The learning effect is associated with performance improvements that follow from practice; the memory effect is associated with performance improvements that follow from recognizing specific sentences. The sentence lists were presented to a group of HI listeners in three conditions: (i) Unknown lists presented at the first visit (“first visit test"); these sentences were not affected by any learning or memory effects; (ii) unknown lists presented at the second visit (“second visit test"); these sentences were affected by learning effects, but not by memory effects; (iii) the lists from the first visit presented again at the second visit (“second visit retest"); these sentences were affected by both learning and memory effects. The pure learning effect can thus be estimated from the SRT_N difference between “first visit test" and “second visit test". The pure memory effect can be estimated from the SRT_N difference between “second visit test" and “second visit retest".

3.5.1 Method

Listeners

Twelve (9 male, 3 female) HI listeners participated. Their age was between 59 and 72, mean 64.8 years. The requirements for the listeners in this group were the same as for the previous HI group (although the age requirement was slightly violated for three listeners).

Apparatus and procedure

The test procedure was similar to that of the test validation with NH and HI listeners, but the present group was only tested with a subset of five test lists at the first visit. The 10 test lists were counterbalanced across listeners, so each list was included in half (six) of the subsets. The order of the lists was also counterbalanced to avoid order effects.

The retest was conducted three weeks later with the same individual noise levels as during the first visit. For one listener, the time between test and retest was 5.5 weeks. The order of the “old" lists was the same as during the first visit. The order of the “new" lists was counterbalanced across listeners. "Old" and "new" lists interleaved through the session.

3.5.2 Results

Three mean SRT_Ns were calculated for each listener: (i) The mean SRT_N across the five lists presented at the first visit; (ii) the mean SRT_N across the five lists presented for the first time at the second visit; (iii) the mean SRT_N across the five lists presented for the second time during the second visit. For each listener, the means were normalized with respect to the mean SRT_N of the “second visit test" in order to remove the large SRT_N differences between listeners. The results are shown in Fig. 3.11. The mean SRT_Ns across listeners in the three conditions were: 0.10 dB for “first visit test"; 0 dB for “second visit test"; and -0.15 dB for “second visit retest". The combined training effect (learning and memory) was thereby -0.25 dB from test

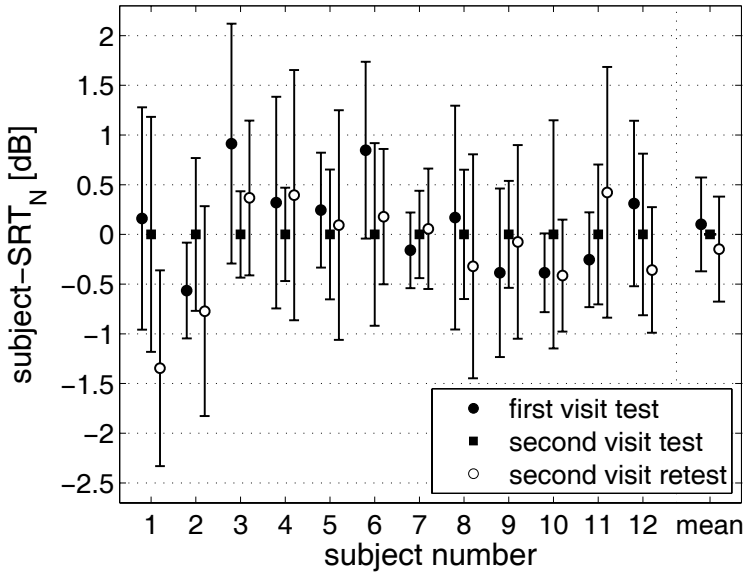


Figure 3.11: Comparison of the subject-SRT$_N$s determined across five lists in three conditions. The bars indicate ± 1 within-subject standard deviation. The “mean” entry shows the overall mean in the three conditions with bars indicating the standard deviation of the subject means. Values are normalized with respect to the SRT$_N$ of the “second visit test". A decrease of the SRT$_N$ from the “first visit test" to the “second visit test" is an indication of a learning effect. A decrease of the SRT$_N$ from the “second visit test" to the “second visit retest" is an indication of a memory effect.

to retest. The effect was slightly lower than for the previous group of HI listeners (-0.36 dB).

3.5.3 Discussion

The differences in SRT$_N$ between the three test conditions for this group of HI listeners were small. The estimate for the pure learning effect after three weeks is -0.10 dB. The pure memory effect is estimated to be -0.15 dB. However, this estimate is dominated by a particularly low retest SRT$_N$ for listener 1. Removing listener 1 from the calculation reduces the memory effect to -0.04 dB.

The results from this group confirm the results for the NH listeners and the previous HI listeners (Fig. 3.5) that the SRT_N change between test and retest is below 0.5 dB for most listeners. The results also imply that only half or probably less of the subject-SRT_N decrease between test and retest is due to a memory effect.

3.6 Conclusion

The present test validation with NH listeners established normative data for the test that are comparable to those of HINT in other language versions (Soli and Wong, 2008). The normative SRT_N of -2.52 dB for the Danish HINT is slightly higher than for other HINTs, and it is substantially higher than the value obtained with another Danish speech test, the DANTALE II (with a normative SRT_N of -8.4 dB). The normative standard deviation of the SRT_N for the Danish HINT is slightly below the mean for the HINTs reported in Soli and Wong (2008).

The validation with HI listeners led to SRT_N assessments with a within-subject deviation and a between-list deviation that were only slightly different from those obtained with NH listeners. The test is thus expected to produce results that are equally reliable for NH and HI listeners.

The test and retest with a three week interval showed only small differences in the measured SRT_Ns. Changes in the subject-SRT_Ns were generally below 0.5 dB. Reuse of the test lists after three weeks thus seems applicable. The investigation of the separated learning and memory effects suggest that recollection of the sentences only accounted for a minor part of the SRT_N decrease between test and retest.

Acknowledgements

This study was conducted in collaboration with the Danish hearing aid companies Oticon, GN Resound, and Widex. We wish to thank the following colleagues for their contributions and involvement: Lise Bruun Hansen and Niels Søgaard Jensen (Oticon); Anja Kofoed Pedersen and Ellen Raben Pedersen (Widex); Charlotte T. Jespersen, Jenny Nesgaard, and Lotte Hernvig (GN Resound). We would also like to thank Julie Neel Weile for her testing of the many HI listeners. Finally, thank you

to all the listeners who took time to participate. The present work was partly funded by the Oticon Foundation.

Chapter 4

Revisiting extrinsic compensation for reverberation

This chapter is based on Nielsen and Dau (2009c)

Abstract

Listeners were given the task to identify the stop-consonant [t] in the test-word “stir” when the word was embedded in a carrier sentence. Reverberation was added to the test-word, but not to the carrier, and the ability to identify the [t] decreased because the amplitude modulations associated with the [t] were smeared. When a similar amount of reverberation was also added to the carrier sentence, the listeners’ ability to identify the stop-consonant was restored. This phenomenon has in previous research been considered as evidence for an extrinsic compensation mechanism for reverberation in the human auditory system [Watkins, J. Acoust. Soc. Am. 118, 249-262 (2005)]. In the present study, the reverberant test-word was embedded in additional non-reverberant carriers, such as white noise, speech-shaped noise and amplitude modulated noise. In addition, a reference condition was included where the test-word was presented in isolation, i.e., without any carrier stimulus. In all of these conditions, the ability to identify the stop-consonant [t] was enhanced relative to the condition using the non-reverberant speech carrier. The results suggest that the non-reverberant speech carrier produces an unintended interference that impedes the identification of the stop-consonant. These findings raise doubts about the existence of the compensation mechanism.

4.1 Introduction

Reverberation can have a significant impact on the intelligibility of speech. Apart from background noise, it is the environmental factor most often responsible for a poor intelligibility. Reverberation smears the amplitude modulations of a sound signal, so energy peaks are prolonged and become less pronounced while energy dips are masked by preceding sounds (e.g., Houtgast and Steeneken, 1985). The attenuation of amplitude modulations in a room is an accurate predictor of speech intelligibility, and this correlation is reflected in methods for assessing intelligibility, such as the speech transmission index (STI; Steeneken and Houtgast, 1980). In a series of studies (e.g., Watkins, 2005b,c), Watkins considered the reduction of amplitude modulations in investigations of the impact of reverberation on speech intelligibility. He tested the ability of listeners to identify the stop-consonant [t] in the test-word "stir" when the word was embedded in a carrier sentence and different amounts of reverberation were added. When listeners could not identify the [t], the test-word was perceived as "sir". Stop-consonants are particularly susceptible to reverberation (Helfer, 1994) because the amplitude modulation dip that is an important feature for their identification is attenuated. Drullman et al. (1994) noted that stop-consonants are sensitive to reverberation because their identification depends on rapidly changing amplitudes. Rather than focusing on the negative influence of reverberation, Watkins proposed that the human auditory system can, to a large extent, *compensate* for the effects of reverberation. His experiments demonstrated that when reverberation was added to the test-word, but not to the carrier sentence, more words were identified as "sir" because the modulations associated with the [t] in "stir" were smeared. However, when a similar amount of reverberation was also added to the carrier sentence, the number of "stir" identifications increased again. Watkins referred to this effect as *perceptual compensation* because he assumed that it was based on an enhanced ability of the listener to perceptually differentiate between the direct sound and the reverberation. Watkins also characterized the effect as *extrinsic* because it depends on information about the reverberation from the surrounding speech carrier. When the reverberation of the carrier is similar to the reverberation of the test-word, the compensation mechanism enhances the intelligibility of this word.

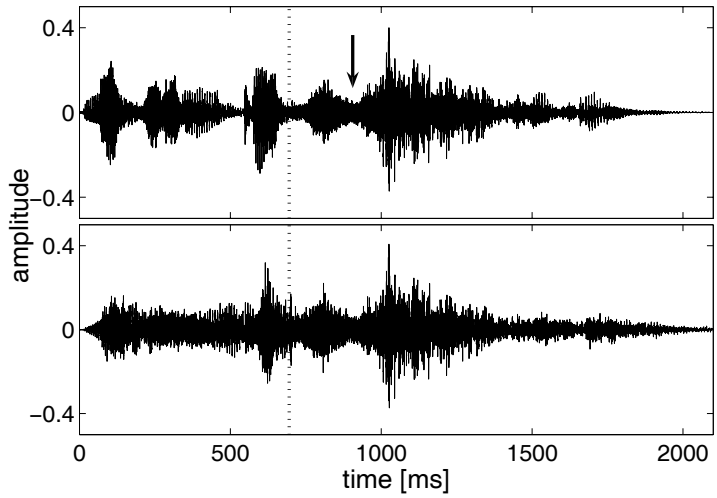


Figure 4.1: Two examples of stimuli used by Watkins (2005c). Both plots show the sentence "next, you'll get..stir..to click on". The vertical line marks the divide between the preceding carrier and the test-word. The arrow points to the amplitude modulation dip originating from the [t] in "stir". In the upper plot, the carrier is slightly reverberant while the test-word is strongly reverberant. Below, the reverberation of the carrier is raised to the same level as for the test-word. The amplitude modulations of the carrier in the upper panel are more pronounced than in the lower panel.

The key observation in Watkins' investigations of the compensation mechanism was the shift in the listeners' perception of the reverberant test-word when embedded in two different speech carriers: (i) an almost non-reverberant carrier, and (ii) a carrier with a reverberation that matches the reverberation of the test-word. The interpretation of this shift as evidence for extrinsic compensation requires that the change in carrier reverberation *per se* causes the shift. However, adding reverberation to a speech signal changes several of the acoustical properties of the signal. Modulation depths are reduced and transitions used for phonetic identification are smeared. Furthermore, by adding different amounts of reverberation to the different words of the stimulus sentence, a signal with varying acoustic-phonetic characteristics is created. Several studies have shown that the perception of a word is affected when it is embedded in sentences with different acoustic-phonetic characteristics (e.g., Fourcin, 1968; Mullennix et al., 1989). Acoustic-phonetic variability in a speech signal can significantly

impair both the identification of, and memory for, spoken words (Sommers and Barcroft, 2006). The shift in the perception of the "sir/stir" test-word when reverberation is added to the carrier might then not be the result of a compensating effect produced by the reverberant carrier, but the result of removing the acoustic-phonetic variability that was present in the combined non-reverberant/reverberant stimuli. In combination with the reverberant test-word, the non-reverberant carrier might then be regarded as producing an interfering effect that impedes the identification of the [t] in "stir". Such an alternative interpretation of the data has not been considered in the studies by Watkins.

Modulation adaptation, also sometimes referred to as forward masking in the modulation domain (e.g., Wojtczak and Viemeister, 2005), could be another reason why the perception of the [t] in "stir" changes when reverberation is added to the carrier. In Fig. 4.1, Watkins' original carrier sentence, "next, you'll get... to click on", is shown with the test-word "stir" embedded. In both panels, the part comprising "stir" is highly reverberant, but the distinctive modulation dip associated with the stop-consonant [t] is still visible (indicated by the arrow). In the upper panel, the carrier is essentially non-reverberant whereas carrier and test-word reverberations are matched in the lower panel. The amplitude modulations of the non-reverberant carrier (upper panel) are clearly more pronounced than those of the reverberant carrier (lower panel). Amplitude modulations in a stimulus can mask modulations of similar rate in a subsequently presented stimulus and the amount of masking increases with increasing modulation depth of the masking stimulus (e.g., Wojtczak and Viemeister, 2005). The ability to identify the [t] in "stir" might thus decrease when the test-word is preceded by the more strongly modulated (non-reverberant) carrier than by the less modulated (reverberant) carrier. As a consequence, a listener might make more "stir" judgements with the reverberant carrier because the modulation depth of this carrier is reduced compared to the non-reverberant carrier.

The hypothesis of the present study is that an extrinsic compensation mechanism for reverberation does not exist. Instead, it is proposed that the shift in the perception of the [t] in "stir" when switching between a non-reverberant and a reverberant speech carrier is caused by an interfering effect stemming from the non-reverberant carrier. The effect is assumed to be caused by acoustic-phonetic variability and/or forward

modulation masking when stimuli with different levels of reverberation are combined. Two experiments were conducted to test this. The purpose of the first experiment was to confirm Watkins' observations of a perceptual shift of the test-word when changing the reverberation of the carrier (Watkins, 2005c). This experiment also included a condition where the test-word was embedded in a non-reverberant white-noise carrier. If the non-reverberant noise carrier and the reverberant speech carrier would produce similar effects on the perception of the test-word, this would indicate that the shift in perception was not caused by compensation for reverberation, but rather by other effects. In the second experiment, additional conditions with non-reverberant carriers were tested, including a speech-shaped noise, an amplitude-modulated noise, and a condition with a silent carrier, i.e., a pause, in order to further examine which carrier characteristics affect the perception of the test-word.

4.2 Methods

4.2.1 Experimental procedure

The same experimental procedure as in Watkins (2005c) was used. In that study, binaural room impulse responses (BRIRs) were recorded at different distances from a sound source; the extreme distances were 0.32 m (referred to as "near") and 10 m (referred to as "far"). The BRIRs were convolved with dry speech recordings in order to achieve test-word and carrier stimuli with different levels of reverberation. The influence of reverberation on the identification of the stop-consonant [t] was assessed using a continuum of 11 words, changing in steps from plain "sir" (step 0) to plain "stir" (step 10). The words were generated by determining the envelopes of "sir" and "stir" and imposing different ratios of these envelopes on the waveform of "sir" (refer to Watkins, 2005c, for details). Test-words corresponding to low step numbers were perceived as "sir" and words corresponding to high step numbers were perceived as "stir". The 11 words were presented in combination with different carriers to the listeners. All stimulus combinations were repeated three times. The listeners switched from identifying the test-word as "sir" to "stir" at the so-called *category boundary*. This transition from "sir" to "stir" typically took place within a few continuum steps.

The category boundary quantified the listeners' ability to identify the [t] in the test-word and was calculated as the total number of "sir" responses divided by 3 minus 0.5 (as in Watkins, 2005c).

4.2.2 Speech stimuli

The speech stimuli were based on Watkins' original carrier ("next you'll get ... to click on") and his original "sir" to "stir" continuum. The original sound files consisted of a carrier and an embedded test-word in three combinations of reverberation: (i) "near" carrier and "near" test-word; (ii) "near" carrier and "far" test-word; (iii) "far" carrier and "far" test-word. For each combination, the carrier was combined with all 11 test-words from the continuum (from plain "sir" to plain "stir") resulting in 33 different sound files. The "near" reverberation produced a sound that was relatively dry with hardly noticeable reflections. With the "far" reverberation, the reflections were clearly noticeable, but the intelligibility of clear speech was not affected. Watkins used fast as well as slowly spoken stimuli in his investigations; the stimuli used here were part of the slow version.

The experiments in the present study required that the test-words were combined with other carriers than the speech carrier of Watkins (2005c). The original stimuli were therefore separated into individual waveforms. First, the 11 "near" sound files were divided into a "near" start carrier waveform ("next you'll get ..."), a "near" end carrier waveform ("...to click on") and 11 separate "near" test-word waveforms. The reverberation time of the "near" waveforms was sufficiently short for the reverberation "tails" to be inaudible in the subsequent and now separated part of the signal. Second, one of the original "far" version sound files was deconvolved with a corresponding "near" version to derive an estimate of the "impulse response" of the room. Third, the separated "near" waveforms were convolved with this impulse response to obtain corresponding "far" versions of the waveforms. The available stimuli now consisted of separate "near" and "far" versions of the start carrier, the end carrier, and all 11 continuum steps of the test-word.

The reverberation tails of the convolved carriers and test-words contained a small amount of ringing. The convolved waveforms were concatenated in order to make a comparison with the original "far" sound files. The concatenations were time-

aligned with the original stimuli with respect to the onsets of the start carrier, the test-word, and the end carrier. Informal listening tests revealed that the original and the concatenated stimuli were hardly distinguishable.

4.2.3 Apparatus and procedure

The experiments were conducted in a double-walled sound insulated booth with a screen, a keyboard and a mouse connected to an external PC. The stimuli were presented over Sennheiser HD580 headphones at an average sound pressure level of 65 dB. A PC-application played the stimuli and waited for the listener to respond either “sir” or “stir” by a mouse click on one of two clearly marked buttons on the PC screen. The listener could alternatively respond by pressing “1” or “2” on the keyboard. After the response from the listener, a 2 s silent interval was presented before the next stimulus. The experiments were preceded by a 1-2 min. training sequence.

4.3 Experiment 1: “Sir” versus “stir” identifications using the original setup

4.3.1 Rationale

This experiment was performed to reproduce the key finding of Watkins (2005c) that the number of “stir” identifications increases when the amount of reverberation added to the carrier is increased. In addition, a white noise that did not contain any reverberation information was included as a carrier. In terms of compensation for reverberation, this carrier should have a negligible effect on the listener’s perception of the test-word.

4.3.2 Method

Listeners

Six listeners participated in the experiment. They were aged between 24 to 42 and did not report any hearing problems. They were all students or employees of the Technical University of Denmark (DTU) and had previous experience with psychoacoustic experiments. The second author participated in the experiment. The listeners were fluent English speakers. All experiments in this study were approved by the ethics committee of Copenhagen County.

Stimuli

Watkins' 11 step "sir" to "stir" continuum was presented in three combinations with the "next you'll get ... to click on" carrier: (i) "near carrier - near test-word", (ii) "near carrier - far test-word", and (iii) "far carrier - far test-word". The PC application that ran the experiment concatenated the separated carriers and test-words to recreate Watkins' original stimuli. The "far" version of the test-word was also presented in combination with a white-noise carrier. The white noise was presented at the same sound pressure level as the speech carrier (calculated separately for the start and the end portions of the carrier). All stimuli were presented diotically. All combinations were presented three times in random order, resulting in 4 combinations x 11 continuum steps x 3 repetitions = 132 presentations.

4.3.3 Results and discussion

Figure 4.2 shows the individual results of experiment 1 for the six listeners. For three of the carrier/test-word combinations ("near - near", "far - far", and "wn - far"), the transition from "sir" to "stir" typically occured within 2-3 steps, while the transition was less consistent for the "near - far" combination. Nevertheless, the category boundary was calculated for all combinations. For example, the boundary of the "near - far" combination in the upper left panel was calculated as $(4 \cdot 3 + 2 + 3 + 1 + 2 + 3 \cdot 1)/3 - 0.5 = 7.2$, using the procedure in Watkins (2005c). For all listeners, the "near - far" combination results in a transition from "sir" to "stir" at the relatively highest step

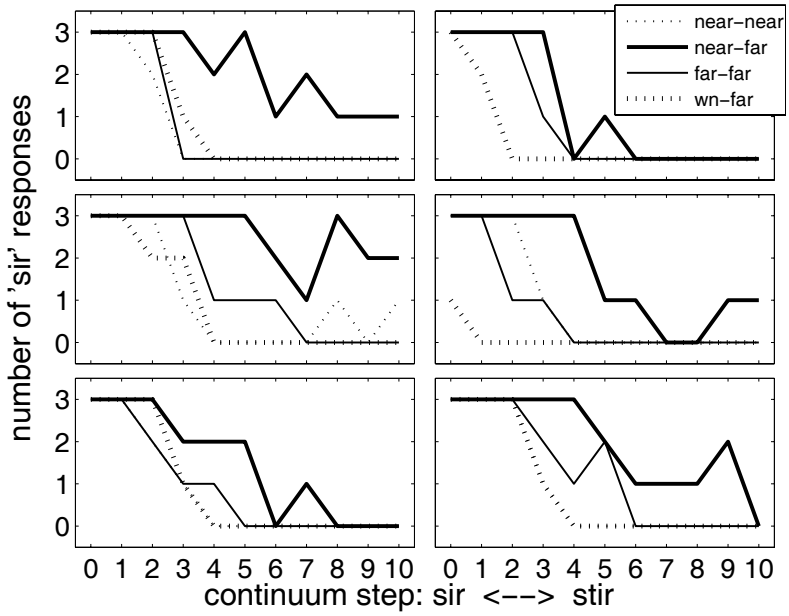


Figure 4.2: Individual results for the six listeners in experiment 1. The abscissa indicates the continuum step of the test-word. The ordinate represents the number of “sir” responses out of three repetitions. For the listener in the upper left panel, the category boundaries are: 2.2 for “near-near”, 2.5 for “far-far”, 2.8 for “wn-far”, and 7.2 for “near-far”.

number, indicating that the identification of the [t] in “stir” was most difficult for this combination. For the other combinations, the functions sometimes cross each other and result in category boundaries with no consistent deviations from each other.

The mean category boundaries across the six listeners are shown in Fig. 4.3. The boundary in the “near - near” condition was found to be at a value of 2.9. Adding reverberation to the test-word (“near - far” condition) shifted the boundary to 6.2, while adding reverberation to the carrier also (“far - far” condition) shifted it back to 3.2, close to the step number of the boundary for the “near - near” condition. The “wn - far” boundary was at an even lower step number of 2.1. Paired t-tests showed that there were no significant differences between the three low-level boundaries, but a highly significant difference between each of these and the “near - far” boundary [$p < 0.007$]. The observed category boundaries are in accordance with the results of

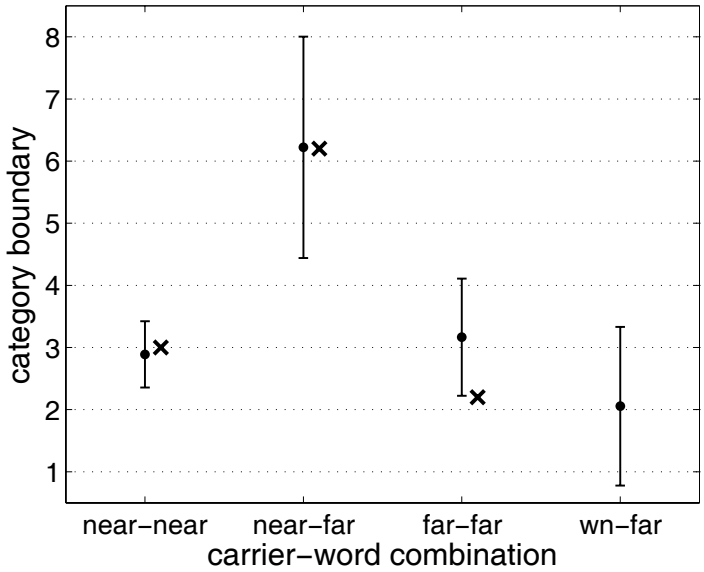


Figure 4.3: Mean category boundaries for the six listeners in experiment 1. Bars indicate one standard deviation. The boundary for the "near - far" condition is significantly higher than for the remaining conditions. The results from a similar experiment in Watkins (2005c) (experiment 1, "slow, L-shaped" condition) are marked by crosses.

experiment 1 for the "slow, L-shaped" condition shown in Fig. 2 of Watkins (2005c). These values are marked by crosses in Fig. 4.3 (mean values only).

Experiment 1 confirms the observations obtained by Watkins (2005c). When added to the test-word, the "far" reverberation significantly shifts the category boundary upward because it becomes more difficult to identify the [t] in "stir". When this reverberation is also added to the carrier, the boundary shifts back, close to the step number in the "near - near" condition. Thus, the carrier reverberation facilitates the identification of the [t], which in principle might be a consequence of the proposed compensation mechanism for reverberation. However, this interpretation appears inconsistent with the low category boundary observed for the "white noise - far" condition. The boundary shift between the "near" and the "white-noise" carrier cannot be caused by a compensation effect related to reverberation since no reverberation was

present in the white noise. Instead, this result indicates that the listeners' perception of the "sir" to "stir" continuum is affected by other carrier characteristics than merely the amount of reverberation.

The effects of a noise carrier were also investigated by Watkins (2005c). He used a noise with the same magnitude spectrum and the same temporal envelope as the original dry speech carrier and convolved the noise with the "near" and "far" BRIRs. His experimental result was similar to that obtained for the noise carrier in the present experiment: The category boundaries for the noise carriers ("near" and "far") were at the same low level as for the "far" speech carrier. This implies that both noise carriers have had a "compensating" effect similar to that of the "far" carrier. However, since the "near" noise only contained a negligible amount of reverberation, it cannot have produced extrinsic compensation for reverberation. Watkins therefore suggested that the boundaries for the noise contexts might be at this low level because an *intrinsic* form of compensation was involved in these conditions. He proposed that such an intrinsic compensation might assess the test-word's reverberation from only the test-word (i.e. ignoring the carrier) when the context is noise (Watkins, 2005c). This implies that the proposed intrinsic compensation effect is operational when the test-word is embedded in the noise carrier, but not when it is embedded in the speech carrier. Watkins mentioned effects associated with perceptual grouping (Bregman, 1990) as an explanation for these observations. Nevertheless, the observations support the interpretation that the "near" speech carrier interferes with the listeners' ability to identify the test-word.

4.4 Experiment 2: Effects of other non-reverberant carriers on "sir" versus "stir" identifications

4.4.1 Rationale

The boundary shift between the "near" speech and the "white noise" carrier observed for the "far" test-word in experiment 1 was not related to reverberation. The second experiment was designed to investigate the boundary shift in greater detail. The experiment included other non-reverberant carriers with special characteristics in

order to examine the effect of these on the perception of the test-word. A “silent” carrier was included as a reference condition.

4.4.2 Listeners

The experiment was conducted with 13 listeners aged 26 to 43, who did not report any hearing problems. They were all students or employees of DTU and had previous experience in psychoacoustic experiments. The second author participated in the experiment. All listeners were fluent English speakers.

4.4.3 Stimuli

Five different carrier conditions were presented in combination with the “far” version of the “sir” to “stir” continuum. The different stimulus waveforms are shown in Fig. 4.4 (with step 10 of the “far” test-word) and include: 1) The first part of the original “near” carrier; 2) the first part of the original “far” carrier; 3) a silent pause of the same duration as the original speech carrier; 4) an unmodulated speech-shaped noise with a magnitude spectrum corresponding to that of the talker of the original speech material; 5) the same speech-shaped noise with imposed amplitude modulations randomly distributed between 4 and 8 Hz. The unmodulated speech-shaped noise was produced by superimposing Watkins’ original “near” speech material 150 times with randomly shifted offsets. This noise was multiplied by an amplitude modulation to produce the modulated noise. The average sound pressure level of the different carriers, except the “silent” carrier, was 65 dB. The portion of the original carrier that followed the test-word (“...to click on”) was omitted in this experiment since it did not have a meaningful correlate for the non-speech carriers. According to Watkins (2005a), this omission should not effect the perception of the test-word. All stimuli were presented diotically. The total number of presentations per listener was 5 carriers x 11 continuum steps x 3 repetitions = 165.

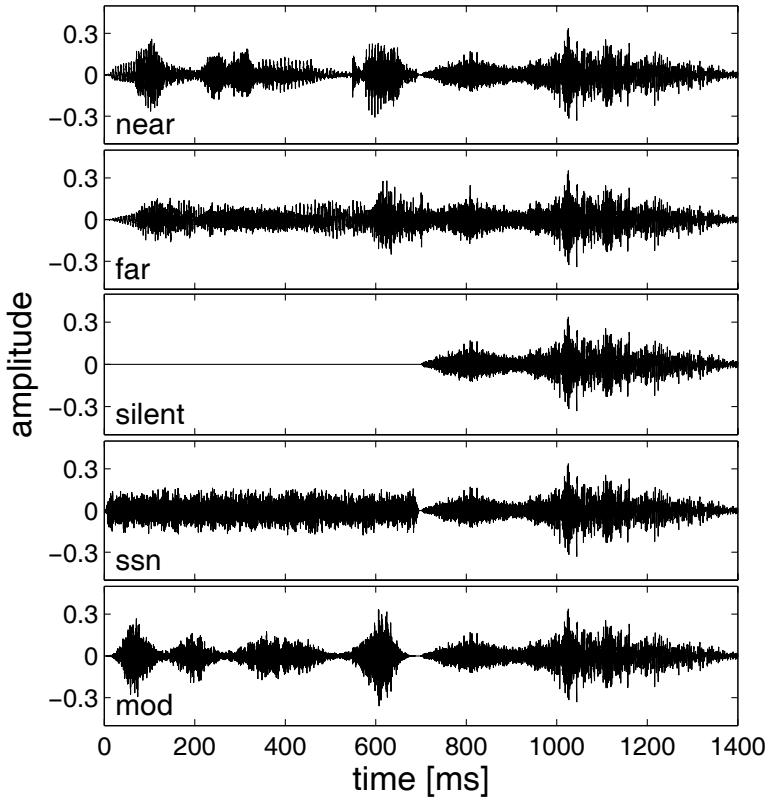


Figure 4.4: The five carriers used as stimuli in experiment 2. As shown, the test-word was not followed by an end carrier in this experiment. All carriers are in this figure combined with the "far" version of step 10 of the test-word continuum.

4.4.4 Results and discussion

Figure 4.5 shows the mean category boundaries for the five different carriers combined with the "far" test-word. The category boundary was 5.3 for the original "near" carrier, 3.7 for the original "far" carrier, 3.7 for the "silent" carrier, 3.3 for the "ssn" carrier, and 4.1 for the "mod" carrier. A paired t-test for the original "near" carrier in combination with the "far", the "silent", and the "ssn" carrier, respectively, showed a significant shift of the boundary [$p < 0.0006$] for all three combinations. There

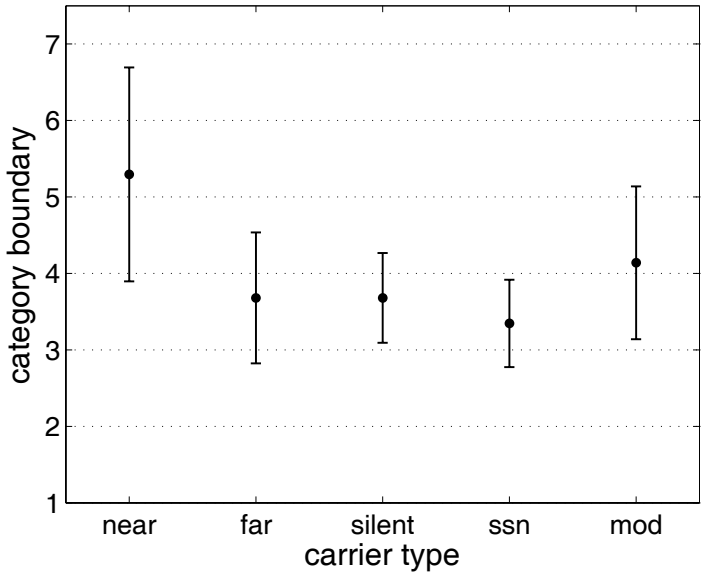


Figure 4.5: Mean category boundaries obtained in experiment 2. Bars indicate one standard deviation. The boundary for the "near" carrier is significantly higher than the boundary for the three middle carriers. The boundary for the modulated noise carrier ("mod") is significantly higher than the boundary for the unmodulated noise ("ssn").

were no significant differences between the "silent", the "ssn", and the "far" carriers. Again, this was not predicted by the hypothesis of a compensation effect related to the reverberation of the "far" carrier. This carrier does not have any particular effect that cannot be obtained with a speech-shaped noise or even a silent interval. The "near" speech carrier is indeed the only one of the four leftmost carriers in Fig. 4.5 with a significantly different boundary level and therefore the only one with an exceptional effect on the listeners' perception of the test-word. The effect might be caused by the acoustic-phonetic variability that is introduced when the "near" carrier is combined with the "far" test-word. In terms of potential underlying mechanisms, the interfering, or masking, effect might be caused by the specific acoustical characteristics of the "near" carrier such as its spectro-temporal modulations. This is, at least partly, supported by the findings obtained with the modulated noise carrier ("mod"), shown

in the rightmost position in Fig. 4.5. The boundary of this carrier was significantly higher than the boundary of the unmodulated noise carrier ("ssn") [$p < 0.03$]. These two carriers differ with respect to their amplitude modulations (Fig. 4.4), but both stimuli were non-reverberant. This suggests that the listeners' perception of the test-word has shifted because of the differences in the amplitude modulation depth of the two carriers. As shown in Fig. 4.1, also the original "near" and "far" speech carrier waveforms differ with respect to their degree of modulation. It is, thus, likely that part of the boundary shift between these two carriers is due to a change in their amplitude modulation content.

4.5 General discussion

4.5.1 Summary of the main findings

The evidence of extrinsic compensation in Watkins (2005c) was based on an increased number of "stir" identifications in a "sir" to "stir" continuum when the reverberant test-word was embedded in a speech carrier with the same amount of reverberation. In the experiments of the present study, a similar increase of "stir" identifications was found with other carriers that did not contain any information related to the reverberation of the test-word. In fact, all other carriers tested in this study led to lower category boundaries than the original "near" speech carrier. This suggests that the boundary shifts between the "near" and the "far" speech carriers in Watkins (2005c) were not caused by a compensation effect. Instead, the results favor the interpretation that the "near" carrier produces an interfering effect on the identification of the [t] in "stir".

The condition using the "silent" carrier can be regarded as a reference condition since the listeners' perception of the test-word was unaffected by any preceding sound for approximately 3 s. There was no difference between the boundary obtained with the "silent" carrier and that obtained with the "far" speech carrier, but a highly significant shift of the boundary between the "silent" carrier and the "near" speech carrier conditions. Thus, if any of the two speech carriers have an exceptional effect

on listeners' perception of the test-word it should be ascribed to the "near" carrier and not to the "far" carrier.

4.5.2 Potential causes of perceptual interference

The acoustic-phonetic variability that is introduced by combining a carrier sentence and a test-word with different levels of reverberation might partly cause the interfering effect of the "near" carrier. Mullennix et al. (1989) reported that trial-to-trial changes of the talker of spoken words reduced the word recognition compared to a situation with no change of the talker. Similarly, Sommers and Barcroft (2006) showed that speaking style variability reduced the identification performance for spoken words compared to the single speaking style condition; this was true for all (six) tested speaking styles. Sommers et al. (1994) suggested that acoustic-phonetic variability will impair spoken word identification when the variability alters acoustic properties that are used for phonetic identification. Phonetic transitions are one of the primary cues for stop-consonants as the [t] in "stir". These phonetic transitions are changed by reverberation and are sharper in Watkins' "near" stimuli than in the "far" stimuli. The identification of the [t]-phoneme in "stir" might therefore be negatively affected by the acoustic-phonetic variability that is introduced when the reverberation switches between the "near" carrier and the "far" test-word.

An alternative interpretation, that is more related to concepts of auditory masking, is that the identification of the [t] phoneme is affected by the modulation content of the preceding carrier. This is supported by the result obtained with the modulated noise carrier, where the boundary was higher than for the unmodulated noise (condition "mod" and "ssn" in Fig. 4.5). Here, the increased amount of modulation energy in the speech relevant modulation frequency range ($\approx 4 - 20$ Hz) may have produced an increased amount of modulation adaptation. Wojtczak and Viemeister (2005) measured forward masking in a modulation detection task and found that the sensitivity to amplitude modulations can be substantially decreased even after a brief exposure to modulations of similar rate. The duration of the (non-speech) masker in Wojtczak and Viemeister's experiments was 150 ms. They observed an exponential recovery from the masking effect and a threshold that remained elevated for at least 150 ms. This is approximately the time interval between the offset of the "near" carrier and

the position of the [t] in “stir”. A forward masking, or modulation adaptation, effect stemming from the relatively deep modulations contained in the “near” carrier might therefore be responsible for part of the interfering effect produced by this carrier.

However, since adaptation effects are highly audio-frequency and modulation-frequency specific (Kay and Matthews, 1972), it is clear that an amplitude modulated broadband noise cannot produce the same amount of adaptation as speech, which consists of complex amplitude and frequency modulations varying over time. Specifically, a carrier containing dry speech of the same speaker can be expected to produce a maximal interference in this particular task. The observation that the amplitude modulated noise carrier produced a significant effect on the identification results shows that the modulation energy contained in the carrier is crucial.

4.5.3 Consistency between the data from the original and the present study

The assumption of an interfering effect produced by the “near” carrier as an alternative explanation for the boundary shifts observed in Watkins (2005c) is supported by some of the original experimental results in that same study.

In experiment 2 of Watkins (2005c), the reverberation that was added to the test-word and the speech carrier, respectively, originated from two different locations, an L-shaped room and a corridor. The results were compared to the original experimental setup with reverberation from the *same* location applied to the carrier and the test-word (Fig. 3 in Watkins, 2005c). It was found that the compensation effect was essentially independent of whether there was a switch in location between the carrier and test-word or not, and the author concluded that the compensation effect is independent of details in the reverberation. Instead, the effect seemed to rely on aspects of the reverberation that were common in the two locations. It is difficult to understand what kind of information a reverberation compensation mechanism can exploit from a carrier, when the origin of the reverberation is without relevance. However, an interpretation of the boundary shift based on an interfering effect produced by the “near” carrier would explain why the switch of location has little importance. The “near” BRIRs of the L-shaped room and the corridor are similar because the direct

sound dominates. The two "near" speech carriers are thus similar and their interfering effect will be approximately the same.

In the same study, Watkins (2005c) also investigated the effect of dichotic versus monaural presentation of the stimuli. The extrinsic compensation for reverberation was found to be *greater* in the monaural condition than in the dichotic condition (Fig. 3 in Watkins, 2005c). This result seems to be inconsistent with the fact that a dichotic signal contains more information about the reverberation in a room than a monaural signal and that binaural listening (relative to monaural listening) typically improves speech intelligibility in reverberant rooms (e.g., Nábělek and Robinson, 1982). The author concluded that extrinsic compensation does not use binaural information, but stems from a monaural mechanism. He also argued that binaural advantages for listening in reverberation are "intrinsic", and thus not affected by changes in the extrinsic carrier reverberation. Watkins hereby established two distinctly different ways for the auditory system to benefit from reverberation information: An analysis of "intrinsic" reverberation based on binaural information and an analysis of "extrinsic" reverberation based on monaural information. This distinction between intrinsic and extrinsic reverberation and the assumption of two different auditory approaches to their analysis is difficult to verify. An alternative interpretation that complies with the hypothesis of an interfering effect caused by the "near" carrier, would be that this effect is essentially monaural, but can be reduced to some extent by binaural auditory processing.

The study of Watkins (2005c) also included measurements of the effect of carrier reverberation on "near" test-words that only contained a very small amount of reverberation. These conditions were not considered in the present study. The results showed that the category boundary obtained for the "near" words depends on the carrier reverberation in the same manner as for the "far" test-words: The boundary was lower when the test-word was combined with the "far" carrier than when combined with the "near" carrier (Figs. 2, 3, and 4 in Watkins, 2005c). There was a floor-effect restricting the size of the boundary shift for the "near" test-word, nevertheless, the same direction of the shift was observable in all experiments. However, this result does not seem compatible with the concept of reverberation compensation since extrinsic compensation should result in more "stir" identifications when the

"reverberation information" of the carrier matches the reverberation information of the test-word. Instead, Watkins' data showed that the "far" carrier *always* led to more "stir" identifications, even when the carrier was misleading with respect to the reverberation of the test-word. This further suggests that the shifts of the category boundary cannot be related to compensation for reverberation, but must be related to other carrier characteristics.

While the results from the present study seem difficult to explain in terms of compensation for reverberation, it remains unclear which specific auditory mechanisms actually cause the boundary shifts. The results of the present study only suggest that the higher values for the boundary were caused by an interference produced by the specific spectro-temporal properties (such as amplitude and frequency modulations) of the carrier preceding the test-word. However, further experimental work as well as auditory modeling are needed to obtain a better understanding of the underlying mechanisms.

4.6 Conclusion

This study investigated the compensation for reverberation hypothesis proposed by Watkins (2005c), which states that the human auditory system can perceptually compensate for the negative effect of reverberation on speech intelligibility. The compensation mechanism is assumed to require reverberation information from preceding speech to be operational and has therefore been termed "extrinsic compensation".

The results of the present study are difficult to explain within the concept of this compensation mechanism since most conditions with non-reverberant carriers, including a reference condition containing a silent interval, produced the same results as the reverberant speech carrier. Rather, the results seem consistent with an interfering effect produced by the specific stimulus characteristics of the "near" speech carrier. This carrier produced a category boundary that was distinctly higher than any other tested carrier. It is suggested that the interference effect is, at least partly, related to modulation adaptation effects and the acoustic-phonetic variability that is introduced by switching between a non-reverberant and a reverberant speech signal. Also the original data of Watkins (2005c) seem to be consistent with this interpretation.

Additional studies need to be undertaken, ideally in connection with auditory modeling, to clarify the auditory mechanisms underlying consonant identification in various acoustical contexts.

Acknowledgements

We wish to thank students and colleagues at the Centre for Applied Hearing Research for participation in the experiments, and A. J. Watkins for supplying his original speech stimuli. The present work was supported by the Oticon Foundation.

Chapter 5

Overall discussion

This thesis has investigated and developed methods for measuring speech intelligibility. Reliable assessments of speech intelligibility are essential for investigating speech perception and how it is affected by environmental and human factors. In chapters 2 and 3, the development of two sentence tests, CLUE and the Danish version of HINT, was described. In chapter 4, a method for measuring the intelligibility of a short test-word was investigated. This method was used for studying the impact of reverberation on speech intelligibility.

The development of the first test, CLUE, was aimed at creating a test that resembles the original HINT (Nilsson et al., 1994) as closely as possible. However, during the project, changes in some of the procedures were considered appropriate. The major of these changes was related to the procedure that HINT employs for equalizing the sentence intelligibilities. The HINT procedure is based on an equalization of the average *word* intelligibilities, although an equalization of these does not guarantee equal sentence intelligibilities (see appendix A). As a result, most language versions of HINT are probably based on sentences with intelligibilities that vary more than normally assumed. An investigation of this issue would be interesting. The question would be whether the typical HINT speech material shows large intelligibility deviations between the individual sentences as anticipated in this thesis. Or does the general quality of the sentences with respect to naturalness, talker, etc. ensure an approximately equal intelligibility in spite of the word equalization procedure?

The alternative method for equalizing sentence intelligibilities was shown to produce a higher test list equivalence in CLUE than in other language versions of HINT. This suggests that the CLUE sentences are more homogenous than the

HINT sentences with respect to intelligibility. The CLUE equalization procedure should therefore be considered an option in future sentence test development projects, especially, if an investigation has confirmed that the HINT sentences tend to have large deviations.

However, the CLUE equalization procedure could be further improved. As described in chapter 2, the subjective equalization procedure was conducted without the presence of a test leader. The listeners therefore adjusted the level of each sentence based on a perceived wording that was not checked for deviations from the actual sentence. An example is the sentence "Jeg trykker på knappen igen" (I press the button again). Apparently, the listeners adjusted the RMS level of this sentence without consideration for the last word, so other listeners at later development stages would only hear "Jeg trykker på knappen". The sentence consequently had to be discarded at a late stage of the test development. Future use of the CLUE equalization procedure should therefore include a screening of all sentences before the actual equalization in order to ensure a close accordance between perceived and actual wording.

Compared to the original HINT, the calculation procedure of the SRT_N was slightly changed in CLUE. The last eight, instead of seven, sentence levels were included in the calculation. Considering the current HINT test procedure that is based on 20-sentence lists, the significance of this change may seem rather limited. However, HINTs with 10-sentence lists may still be used in some contexts and new tests with 10-sentence lists may be developed. For such tests, it is still relevant to consider how many sentence levels to include in the SRT_N calculation in order to obtain the most reliable result.

As described in chapter 3, the CLUE test was converted into a test that follows the current HINT standard. The validation of the test with 16 NH listeners led to an SRT_N with a standard deviation of 0.87 dB across lists and listeners. This is a slight improvement compared to the overall standard deviation of CLUE of 1.0 dB. The improvement is, however, minor taking the length of the test lists into consideration. The HINT uses 20 sentences per SRT_N assessment, twice as many as in CLUE. The difference in scoring rules, listeners, etc. may explain why a larger improvement was not observed. There may also be a more fundamental explanation: An SRT_N assessment based on one 20-sentence list might not be as reliable as an assessment

based on the mean of two 10-sentence list measurements. This seems confirmed by the standard deviations for the 20-sentence lists reported in Soli and Wong (2008) (noise front). The average standard deviation is only slightly below the typical standard deviation in a test with 10-sentence lists (e.g., Hällgren et al., 2006; Myhrum and Moen, 2008). 20 sentences per list thereby appears to be a less than optimal use of the available sentence material when constructing a sentence test. The primary reason for using this number of sentences also seems to be historical: 20-sentence lists can easily be constructed by combining two 10-sentence lists, the length of the lists in the original HINT. It could well be that assessments with, say, 15-sentence lists are as reliable as assessments with 20-sentence lists. An investigation of this issue would be highly appropriate before developing a new sentence test.

The validation with HI listeners with mild to moderate hearing loss showed that the Danish HINT is as reliable for this group as for NH listeners. This is a valuable result because it removes an uncertainty that otherwise could be difficult to avoid. Another important result is the SRT_N deviation between test and retest after three weeks. The training effect (the combined effect of learning and memory effects) is so small that the test without problems can be reused with the same listeners after some time. This increases the usability of the test. The investigation of a separate learning and memory effect indicated that the memory effect (the decrease of the SRT_N caused by listeners' ability to remember specific sentences) accounts for half or less of the test-retest variance. With such a small memory effect, the listeners seem to have had only a faint recollection of the sentences at the time of the retest. Therefore, the test lists may even be usable several times with the same listeners without the measurements being affected by memory. The randomization of the sentences within each test list, resulting in a different sentence order at test and retest, has probably contributed positively to a reduction of the memory effect.

The experimental result that the Danish HINT works equally well with NH and HI listeners is probably also valid for the CLUE test, since the speech material, the background noise, and the adaptive test procedure are similar for the two tests. This is consistent with the CLUE evaluation that was done by one of the collaborating hearing aid companies. In a test with a subset of the CLUE lists, the standard deviations for five HI listeners were found to be larger than for two NH listeners, but nevertheless

acceptable (personal communication). In some situations, CLUE may thus still be an alternative to the Danish HINT. This could be in investigations where more than 10 SRT_N measurements per listener are needed within the same test session, or where the time gained by only using 10-sentence lists is important.

Sentence tests like CLUE and HINT can also be used for measuring the effects of reverberation on speech intelligibility by convolving the speech material with binaural room impulse responses (BRIRs). However, this option was not chosen for the investigations of reverberation in chapter 4 because it does not permit the observation of how different reverberant or non-reverberant speech *contexts* affect the intelligibility. For this purpose, Watkins (2005c) developed an alternative method based on a listener's perception of a the stop-consonant [t] in a short test-word. In chapter 4, the main result of Watkins (2005c) was reproduced, but another interpretation was presented to account for the data. This interpretation claims that the listener's perception of the [t]-phone is not primarily affected by the changes in the carrier reverberation *per se* but by the additional changes of the carrier's acoustic-phonetic properties that inevitably follow. The test method thereby demonstrates a common problem in psycho-acoustic measurements. Often it is almost impossible to change one property of a stimulus without introducing cues or artifacts that will help or disturb the listener in an unintended manner.

The difference between the measurement method of sentence tests like CLUE and HINT and Watkins' method raises another question about the validity of the latter. In CLUE and the Danish HINT, the sentence lists are phonetically balanced and practically all Danish phonemes are present in each list. This is considered important for a reliable assessment of the SRT_N. With this in mind, it seems debatable whether a listener's perception of a single [t]-phone is a reliable predictor of speech intelligibility. Even if there were valid evidence of an extrinsic compensation mechanism that enhances the identification of the [t], this would hardly be sufficient as evidence for a mechanism that enhances speech intelligibility in general.

Bibliography

Bio-logic Systems Corp. (2005). *HINT Pro: Hearing in Noise Test User's and Service Manual* (G ed.). Mundelein, Illinois: Bio-logic Systems Corp.

Boothroyd, A. and Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *J Acoust Soc Am*, **84**(1), 101–114.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, **95**(2), 1053–64.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test. *J Acoust Soc Am*, **30**(7), 596–600.

Fourcin, A. J. (1968). Speech source interference. *IEEE Trans. Audio Electroacoust.*, **AU-16**, 65–67.

Glasberg, B. R. and Moore, B. C. J. (1989). Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. *Scand Audiol*, **Suppl. 32**, 1–25.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scand Audiol*, **11**, 79–87.

Hällgren, M., Larsby, B., and Arlinger, S. (2006). A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *Int J Audiol*, **45**, 227–237.

Helfer, K. S. (1994). Binaural cues and consonant perception in reverberation and noise. *J Speech Hear Res*, **37**, 429–438.

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *J Acoust Soc Am*, **37**(1), 158–166.

Houtgast, T. and Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am*, **77**(3), 1069–1077.

Kalikow, D. N., Stevens, K. N., and Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J Acoust Soc Am*, **61**(5), 1337–1351.

Kay, R. H. and Matthews, D. R. (1972). On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. *J Physiol*, **225**, 657–677.

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**(4), 2412–2421.

Luts, H., Boon, E., Wable, J., and Wouters, J. (2008). FIST: A French sentence test for speech intelligibility in noise. *Int J Audiol*, **47**, 373–374.

MacLeod, A. and Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br J Audiol*, **24**, 29–43.

McArdle, R. A. and Wilson, R. H. (2006). Homogeneity of the 18 QuickSIN Lists. *J Am Acad Audiol*, **17**, 157–167.

Middelweerd, M. J., Festen, J. M., and Plomp, R. (1990). Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram. *Audiology*, **29**, 1–7.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *J Acoust Soc Am*, **85**(1), 365–378.

Myhrum, M. and Moen, I. (2008). The Norwegian Hearing in Noise Test. *Int J Audiol*, **47**, 377–378.

Nábělek, A. K. and Robinson, P. K. (1982). Monaural and binaural speech perception in reverberation for listeners of various ages. *J Acoust Soc Am*, **71**(5), 1242–1248.

Nielsen, J. B. and Dau, T. (2009a). The Danish Hearing in Noise Test. *Int J Audiol*. (submitted).

Nielsen, J. B. and Dau, T. (2009b). Development of a Danish speech intelligibility test. *Int J Audiol*, **48**, 729–741.

Nielsen, J. B. and Dau, T. (2009c). Revisiting perceptual compensation for effects of reverberation on speech identification. *J Acoust Soc Am*. (submitted).

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am*, **95**(2), 1085–1099.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am*, **63**(2), 533–549.

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

Prosser, S., Turrini, M., and Arslan, E. (1991). Effects of different noises on speech discrimination by the elderly. *Acta Oto-Laryngol (Stockholm)*, **Suppl. 476**, 136–142.

Soli, S. D. and Wong, L. L. N. (2008). Assessment of speech intelligibility in noise with the Hearing in Noise Test. *Int J Audiol*, **47**, 356–361.

Sommers, M. S. and Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J Acoust Soc Am*, **119**(4), 2406–2416.

Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *J Acoust Soc Am*, **96**(3), 1314–1324.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J Acoust Soc Am*, **67**(1), 318–326.

Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S. D., and Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *Int J Audiol*, **44**, 358–369.

Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am*, **107**(3), 1671–1684.

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.

Watkins, A. J. (2005a). Listening in real-room reverberation: Effects of extrinsic context. In D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Eds.), *Auditory Signal Processing: Physiology, Psychoacoustics, and Models* (pp. 423–428). New York: Springer.

Watkins, A. J. (2005b). Perceptual compensation for effects of echo and of reverberation on speech identification. *Acta Acustica united with Acustica*, **91**, 892–901.

Watkins, A. J. (2005c). Perceptual compensation for effects of reverberation in speech identification. *J Acoust Soc Am*, **118**(1), 249–262.

Wojtczak, M. and Viemeister, N. F. (2005). Forward masking of amplitude modulation: Basic characteristics. *J Acoust Soc Am*, **118**(5), 3198–3210.

Wong, L. L. N. and Soli, S. D. (2005). Development of the Cantonese Hearing In Noise Test (CHINT). *Ear & Hearing*, **26**(3), 276–289.

Appendix A

Comparison of word and sentence intelligibility

Sentence redundancy must be taken into account when comparing the sentence intelligibility (SI) and the average word intelligibility (WI) for a sentence. Boothroyd and Nittrouer (1988) introduced the j-factor, which is equal to the number of independently recognized parts in a sentence. Each part will consist of one or more words. In a normal, conversational five-word sentence, j will be in the order of two to three (Boothroyd and Nittrouer, 1988). The intelligibility of a sentence can be calculated as the product of the part intelligibilities. The examples below show that although sentences are equalized to have the same WI, they cannot be assumed to have the same SI. The examples are based on a target WI of 70% because this corresponds to an SI of approx. 50% under the assumption of two independent parts (j = 2) with equal intelligibility ($0.7 \cdot 0.7 = 0.49$).

Example 1: Independent parts in a sentence do not always contain the same number of words. In a five-word sentence with j = 2, one part may contain four words, the other the remaining one word. And despite the assumptions that lead to a target WI of 0.7, the intelligibility of these two parts may also differ. The intelligibility may be 0.8 and 0.3, respectively. For such a sentence the WI is equal to the target value 70% ($(4 \cdot 0.8 + 0.3)/5 = 0.7$), but the SI is only 24% ($0.8 \cdot 0.3 = 0.24$).

Example 2: The j-factor is likely to vary between different sentences, and may even vary for the same sentence, depending on the SNR (Boothroyd and Nittrouer, 1988). In a sentence with j = 1.5 and a WI of 70% evenly distributed between all

words, the SI will be 59% $(0.7^{1.5} = 0.59)$. In another sentence with the same WI, but with j = 3, the SI is 34% $(0.7^3 = 0.34)$.

Appendix B

CLUE sentence lists

List 1

1. Vinduet vendte ud mod gaden
2. Han/Hun hoppede op på cyklen
3. Den gamle mand smilede stort
4. I regnbuen ses alle farver
5. De/Vi vil hellere male selv
6. Kampen gik godt i begyndelsen
7. Han/Hun har passet sin træning
8. Hver aften spiser de/vi salat
9. Det ringer ud til frikvarter
10. Hans bukser var meget korte

List 2

1. Stuen skal nok blive hyggelig
2. Døren er næsten aldrig åben
3. En ung pige kommer gående
4. De engelske bøffer var møre
5. Han/Hun kunne køre meget stærkt
6. Sofaen står bagerst i rummet
7. Torsdag var han/hun ikke hjemme
8. Begge fodboldhold klarer sig fint

9. Maden blev serveret til tiden
10. Han/Hun havde let ved hovedregning

List 3

1. Skuret er bygget af brædder
2. Hans mor var heldigvis hjemme
3. Under bogen ligger en tegning
4. Han/Hun rensede skærmen for støv
5. De/Vi skal bo på efterskolen
6. Hendes penge var gået tabt
7. Katten kom listende helt stille
8. Blomster og gaver strømmede ind
9. Hun/Han var i strålende humør
10. Vi/De er en fredelig familie

List 4

1. Pigen var køn og velbegavet
2. Vi/De sad ude i køkkenet
3. Flasken var fyldt med æblesaft
4. Katten spinder i hendes arme
5. De/Vi danser på et diskotek
6. Bageren havde tre slags rugbrød
7. Han/Hun kommer mandag med pakken
8. Trøjen er syet af bomuld
9. Hun/Han var en lille solstråle
10. De/Vi kom kørende i hestevogn

List 5

1. Godt håndværk holder i årevis
2. Min/Din kuglepen skriver med rødt
3. Mødet sluttede efter tre timer
4. Han/Hun ønskede sig en jakke
5. Jeg er ikke længere sulten
6. Han/Hun kan lugte hendes parfume
7. Villaen er ikke blevet solgt
8. Hjælpen nåede frem for sent
9. Vi/De spadserede en tur sammen
10. Han/Hun lagde tasken på bordet

List 6

1. Kurven var fyldt med vasketøj
2. Store bølger slog mod stranden
3. Han/Hun lagde brænde på bålet
4. Folk sidder og taler sammen
5. Hun/Han var bedst til matematik
6. Stemningen i klassen er god
7. Hendes mand havde et værksted
8. De/Vi unge gik i biografen
9. Han/Hun trækker gardinet til side
10. Vi/De ligner hinanden ret meget

List 7

1. Børnene sidder i en rundkreds
2. Gæsterne nyder den gode vin
3. Manden ville løbe en tur
4. De/Vi talte lidt om fremtiden

5. Pladsen var spærret af affald
6. Festen varede til over midnat
7. Bakken er halvtreds meter høj
8. Hun/Han havde ingen frakke på
9. De ønsker sig et sommerhus
10. Begge hold scorede otte mål

List 8

1. Forbruget af papir er stort
2. Mandag vågnede vi/de meget sent
3. Hendes far var ikke hjemme
4. I går kom svalerne hertil
5. Jeg havde cyklet i solskin
6. Skoledrengen drikker et glas mælk
7. Butikken holder et stort udsalg
8. Hun/Han lavede en kop kaffe
9. Nu venter landmændene på regn
10. De/Vi kommer sejlende til byen

List 9

1. Vinderen fik en flot pokal
2. Hunden svømmede væk fra kysten
3. De/Vi sidder længe i tavshed
4. Han/Hun læser med stærke briller
5. Pludselig kom der en lastbil
6. Der var altid åbent tirsdag
7. Mine/Dine venner går i gymnasiet
8. Bogen er skrevet på engelsk
9. Der bor mange mennesker her

10. Hun/Han var taget på arbejde

List 10

1. Toget er meget sjældent fuldt
2. Jeg var også utrolig glad
3. Hans datter vil på højskole
4. I går havde filmen premiere
5. Børnene og de voksne sover
6. En taxa kørte langsomt forbi
7. Bilen er ikke længere ny
8. Kaninen sprang ud gennem hullet
9. Næste deltager var smedens søn
10. Jeg sætter mig nede bagved

List 11

1. Reden er bygget af smågrene
2. Nu mangler vi/de blot tallerkner
3. Han/Hun var verdensmester i svømning
4. De/Vi cykler eller tager bilen
5. Huset lå omme bag torvet
6. Jeg spurgte ikke til prisen
7. De/Vi ankom sidst på formiddagen
8. Hun/Han rider på venindens hest
9. Insekter kan flyve meget langt
10. De/Vi har altid boet hjemme

List 12

1. Mødet skal holdes på skolen

2. Udenfor er det fuldstændig mørkt
3. Hun/Han var omgivet af mennesker
4. Børnene kom hjem ved middagstid
5. Bogen var billig på udsalg
6. Cykler kan lejes mange steder
7. Af og til larmer naboerne
8. De/Vi blev hurtigt gode venner
9. Han/Hun afviste det nye forslag
10. Koden til låsen passer ikke

List 13

1. Blomsterne vokser i små skåle
2. Høsten var allerede i hus
3. Vi/De havde en festlig aften
4. Man/Han skal holde korte pauser
5. De/Vi to venner deler arbejdet
6. Hendes kontor ligger langt væk
7. Din/Min bror er meget utålmodig
8. Bogen er fuld af eksempler
9. Manden skal ringe til hende
10. Jeg går ud på dansegulvet

List 14

1. Lakken skal fjernes fra gulvet
2. Han/Hun købte ikke mange blomster
3. Værelset lå ud til baggården
4. Naboerne var med til middagen
5. Lyskrydset skifter snart til rødt
6. Han/Hun er en flittig musiker

7. Vi/De havde en dejlig weekend
8. Lågen bag dem smækkede i
9. Hendes øjne så trætte ud
10. Vi/De får boller og chokolade

List 15

1. Nu skal maskinerne skiftes ud
2. Snart fylder rapporten ti sider
3. Jeg tager fat i dørhåndtaget
4. Tøjet var gået af mode
5. Her går alle med solbriller
6. Kassedamen så venligt på ham
7. Han/Hun ligger stadig i sengen
8. Eleven skriver en lang rapport
9. Hele byen kom til brylluppet
10. Vi/De så lidt af vejrudsigten

List 16

1. Skuffen kunne ikke lukkes helt
2. Vi/De byggede husene af træ
3. I morgen bliver vejret bedre
4. De/Vi sejlede med en husbåd
5. Han/Hun har aldrig lavet middagsmad
6. Udsigten til skoven var god
7. Motorløb kan være ret farligt
8. Vi/De rister pølser over bålet
9. Manden kom til en benzintank
10. Han/Hun kender alle byens gader

List 17

1. Pigen strikker en rød trøje
2. Vi/De ventede længe i køen
3. Om aftenen var der lejrbål
4. Det kilder lidt i fingeren
5. Hun/Han gik hen til telefonen
6. Vi/De skal bare blive siddende
7. Suppen smagte godt af tomat
8. Huset her er hans barndomshjem
9. Redskaber skal sættes på plads
10. Vejrudsigten lover regn og slud

List 18

1. Om morgenen lagde stormen sig
2. Lyden kommer oppe fra loftet
3. Hun/Han har købt en vinterfrakke
4. I spisestuen var lyset tændt
5. Han/Hun talte til en kollega
6. Bagefter skal vi/de have jordbær
7. Musik giver en god stemning
8. Spillerne troede på sig selv
9. Tapetet var faldet af væggen
10. Hun/Han havde de smukkeste øjne

Practice list 1

1. Pigerne går rundt i haven
2. Alle skal betale samme pris
3. Hendes ansigt er stadig solbrændt
4. Filmen blev straks en succes

5. Jeg kan godt lide jazzmusik
6. Vi/De siger tillykke og skåler
7. Chaufføren ser ind i spejlet
8. Snakken ved bordet var livlig
9. Drys retten med hakket persille
10. De mørke pletter skyldes maling

Practice list 2

1. Alle foredrag er på engelsk
2. Drengen stikker hånden langt frem
3. Han/hun stiller mange svære spørgsmål
4. Kagen skal bages i ovnen
5. Båndet blev revet i stykker
6. Klokken var blevet over midnat
7. Han/hun blev en god skolelærer
8. De/vi fik jordbærkage til dessert
9. Jeg skulle ringe til formanden
10. Hatten passer til min/din tøjstil

Practice list 3

1. De to mænd kender hinanden
2. Båden sejler lidt over elleve
3. Fabrikkens port var ikke lukket
4. Hans søster var blevet klippet
5. Jeg ønsker mig et kæledyr
6. Han/hun taler om sit arbejde
7. Natten bliver klar og kølig
8. Tårnet er ikke særlig højt
9. Jeg glemmer aldrig den musik

10. Hendes tøj var helt gennemblødt

Practice list 4

1. Strømperne var gået i stykker
2. Nu begynder en ny sæson
3. Rejsen varer mindst en uge
4. Lad os bare køre igen
5. Første stop er ved svømmehallen
6. Bussen kan ikke komme frem
7. Udsigten er bedst om sommeren
8. Han/hun er tilfreds med artiklen
9. Flyrejsen varer mindst fem timer
10. Jeg tager solbad på stranden

Practice list 5

1. Kunden er tilfreds med svaret
2. Gymnastik gør mig meget stærk
3. Grisene løber frit på marken
4. Holdet er klar til kampen
5. Du skal børste alle tænder
6. Hendes bror vil være brandmand
7. Nu blomstrer roserne på marken
8. Jeg var glad for bryllupsfesten
9. Drengen blev medlem af klubben
10. Renten var kun fire procent

Practice list 6

1. Det var en god fastelavnsfest

2. Kampen skal spilles på onsdag
3. Filmen er rigtig godt lavet
4. Derhjemme spiser vi/de ikke kød
5. Børnene løber rundt og leger
6. Hun/Han kommer meget i teatret
7. Familien går tur i parken
8. Statuen har ikke noget hoved
9. Hun/Han tog en hurtig beslutning
10. Vi/De snakkede med vores venner

Practice list 7

1. Billetterne bliver sendt til os
2. Ikke langt væk ligger rådhuset
3. Posen her er til grøntsager
4. Han/Hun sluttede som nummer fire
5. Chokoladen var dyr og god
6. Byen ser fantastisk dejlig ud
7. Flyttemænd har tit ømme muskler
8. Vi/De sagde farvel til gæsterne
9. Manden kløede sig på armen
10. Arbejdet er hårdt og krævende

Appendix C

HINT sentence lists

List 1

1. Det var en god fastelavnsfest
2. Kampen skal spilles på onsdag
3. Filmen er rigtig godt lavet
4. Derhjemme spiser vi ikke kød
5. Børnene løber rundt og leger
6. Hun kommer meget i teatret
7. Familien går tur i parken
8. Statuen har ikke noget hoved
9. Hun tog en hurtig beslutning
10. Vi snakkede med vores venner
11. Billetterne bliver sendt til os
12. Bussen kan ikke komme frem
13. Posen her er til grøntsager
14. Han sluttede som nummer fire
15. Chokoladen var dyr og god
16. Byen ser fantastisk dejlig ud
17. Jeg skulle ringe til formanden
18. Vi sagde farvel til gæsterne
19. Bakken er halvtreds meter høj
20. Arbejdet er hårdt og krævende

List 2

1. Reden er bygget af smågrene
2. Jeg ønsker mig et kæledyr
3. Han var verdensmester i svømning
4. De cykler eller tager bilen
5. Huset lå omme bag torvet
6. Jeg spurgte ikke til prisen
7. De ankom sidst på formiddagen
8. Hun rider på venindens hest
9. Insekter kan flyve meget langt
10. De har altid boet hjemme
11. Mødet skal holdes på skolen
12. Udenfor er det fuldstændig mørkt
13. Hun var omgivet af mennesker
14. Børnene kom hjem ved middagstid
15. Snakken ved bordet var livlig
16. Alle foredrag er på engelsk
17. Af og til larmer naboerne
18. De blev hurtigt gode venner
19. Han afviste det nye forslag
20. Koden til låsen passer ikke

List 3

1. Om morgenen lagde stormen sig
2. Lyden kommer oppe fra loftet
3. Hun har købt en vinterfrakke
4. Grisene løber frit på marken
5. Han talte til en kollega
6. Bagefter skal vi have jordbær
7. Musik giver en god stemning

8. Spillerne troede på sig selv
9. Tapetet var faldet af væggen
10. Hun havde de smukkeste øjne
11. Hver aften spiser de salat
12. Mandag vågnede vi meget sent
13. Hendes far var ikke hjemme
14. Han er tilfreds med artiklen
15. Klokken var blevet over midnat
16. Båndet blev revet i stykker
17. Butikken holder et stort udsalg
18. Hun lavede en kop kaffe
19. Nu venter landmændene på regn
20. De kommer sejlende til byen

List 4

1. Pigen strikker en rød trøje
2. Vi ventede længe i køen
3. Om aftenen var der lejrbål
4. Det kilder lidt i fingeren
5. Hun gik hen til telefonen
6. Vi skal bare blive siddende
7. Kunden er tilfreds med svaret
8. Huset her er hans barndomshjem
9. Redskaber skal sættes på plads
10. Vejrudsigten lover regn og slud
11. Godt håndværk holder i årevis
12. Min kuglepen skriver med rødt
13. Mødet sluttede efter tre timer
14. Han ønskede sig en jakke
15. Jeg er ikke længere sulten

16. Han købte ikke mange blomster
17. Villaen er ikke blevet solgt
18. Hjælpen nåede frem for sent
19. Hendes bror vil være brandmand
20. Han lagde tasken på bordet

List 5

1. Børnene sidder i en rundkreds
2. Gæsterne nyder den gode vin
3. Manden ville løbe en tur
4. De talte lidt om fremtiden
5. Pladsen var spærret af affald
6. Festen varede til over midnat
7. Manden kløede sig på armen
8. Hun havde ingen frakke på
9. De ønsker sig et sommerhus
10. Begge hold scorede otte mål
11. Stuen skal nok blive hyggelig
12. Døren er næsten aldrig åben
13. Han blev en god skolelærer
14. De engelske bøffer var møre
15. Han kunne køre meget stærkt
16. Sofaen står bagerst i rummet
17. Torsdag var han ikke hjemme
18. Begge fodboldhold klarer sig fint
19. Maden blev serveret til tiden
20. Han havde let ved hovedregning

List 6

1. Nu skal maskinerne skiftes ud
2. Renten var kun fire procent
3. Jeg tager fat i dørhåndtaget
4. Tøjet var gået af mode
5. Her går alle med solbriller
6. Kassedamen så venligt på ham
7. Han ligger stadig i sengen
8. Eleven skriver en lang rapport
9. Hele byen kom til brylluppet
10. Vi så lidt af vejrudsigten
11. Toget er meget sjældent fuldt
12. Jeg var også utrolig glad
13. Hans datter vil på højskole
14. I går havde filmen premiere
15. Fabrikkens port var ikke lukket
16. Hendes tøj var helt gennemblødt
17. Bilen er ikke længere ny
18. Nu begynder en ny sæson
19. Flyrejsen varer mindst fem timer
20. Jeg sætter mig nede bagved

List 7

1. Lakken skal fjernes fra gulvet
2. Han kan lugte hendes parfume
3. Værelset lå ud til baggården
4. Naboerne var med til middagen
5. Lyskrydset skifter snart til rødt
6. Han er en flittig musiker
7. Vi havde en dejlig weekend

8. Udsigten er bedst om sommeren
9. Hendes øjne så trætte ud
10. Vi får boller og chokolade
11. Skuret er bygget af brædder
12. Hans mor var heldigvis hjemme
13. De to mænd kender hinanden
14. Holdet er klar til kampen
15. De skal bo på efterskolen
16. Hendes penge var gået tabt
17. Alle skal betale samme pris
18. Blomster og gaver strømmede ind
19. Hun var i strålende humør
20. Vi er en fredelig familie

List 8

1. Skuffen kunne ikke lukkes helt
2. Vi byggede husene af træ
3. I morgen bliver vejret bedre
4. Han hoppede op på cyklen
5. Han har aldrig lavet middagsmad
6. Udsigten til skoven var god
7. Motorløb kan være ret farligt
8. Vi rister pølser over bålet
9. Manden kom til en benzintank
10. Han kender alle byens gader
11. Pigen var køn og velbegavet
12. Vi sad ude i køkkenet
13. Flasken var fyldt med æblesaft
14. Rejsen varer mindst en uge
15. De danser på et diskotek

16. Bageren havde tre slags rugbrød
17. Han kommer mandag med pakken
18. Tårnet er ikke særlig højt
19. Hun var en lille solstråle
20. De kom kørende i hestevogn

List 9

1. Strømperne var gået i stykker
2. Høsten var allerede i hus
3. Vi havde en festlig aften
4. Man skal holde korte pauser
5. Han taler om sit arbejde
6. Hendes kontor ligger langt væk
7. Din bror er meget utålmodig
8. Bogen er fuld af eksempler
9. Manden skal ringe til hende
10. Jeg går ud på dansegulvet
11. Vinderen fik en flot pokal
12. Hunden svømmede væk fra kysten
13. Hans søster var blevet klippet
14. Han læser med stærke briller
15. Pludselig kom der en lastbil
16. Der var altid åbent tirsdag
17. Mine venner går i gymnasiet
18. Bogen er skrevet på engelsk
19. Der bor mange mennesker her
20. Hun var taget på arbejde

List 10

1. Kurven var fyldt med vasketøj
2. Første stop er ved svømmehallen
3. Han lagde brænde på bålet
4. Folk sidder og taler sammen
5. Hun var bedst til matematik
6. Stemningen i klassen er god
7. Hendes mand havde et værksted
8. De unge gik i biografen
9. Han trækker gardinet til side
10. Vi ligner hinanden ret meget
11. Vinduet vendte ud mod gaden
12. De sejlede med en husbåd
13. Kagen skal bages i ovnen
14. Båden sejler lidt over elleve
15. De vil hellere male selv
16. Kampen gik godt i begyndelsen
17. Han har passet sin træning
18. Forbruget af papir er stort
19. Det ringer ud til frikvarter
20. Hans bukser var meget korte

Practice list 1

1. Pigerne går rundt i haven
2. Hendes ansigt er stadig solbrændt
3. Filmen blev straks en succes
4. Jeg kan godt lide jazzmusik
5. Vi siger tillykke og skåler
6. Chaufføren ser ind i spejlet
7. Drys retten med hakket persille

8. De mørke pletter skyldes maling
9. Drengen stikker hånden langt frem
10. Han stiller mange svære spørgsmål
11. De fik jordbærkage til dessert
12. Hatten passer til min tøjstil
13. Natten bliver klar og kølig
14. Jeg glemmer aldrig den musik
15. Lad os bare køre igen
16. Jeg tager solbad på stranden
17. Gymnastik gør mig meget stærk
18. Du skal børste alle tænder
19. Nu blomstrer roserne på marken
20. Jeg var glad for bryllupsfesten

Practice list 2

1. Drengen blev medlem af klubben
2. Ikke langt væk ligger rådhuset
3. Flyttemænd har tit ømme muskler
4. Nu mangler vi blot tallerkner
5. Bogen var billig på udsalg
6. Cykler kan lejes mange steder
7. I spisestuen var lyset tændt
8. I går kom svalerne hertil
9. Jeg havde cyklet i solskin
10. Skoledrengen drikker et glas mælk
11. Suppen smagte godt af tomat
12. Vi spadserede en tur sammen
13. En ung pige kommer gående
14. Snart fylder rapporten ti sider
15. Børnene og de voksne sover

16. En taxa kørte langsomt forbi
17. Kaninen sprang ud gennem hullet
18. Næste deltager var smedens søn
19. Lågen bag dem smækkede i
20. Under bogen ligger en tegning

Practice list 3

1. Han rensede skærmen for støv
2. Katten kom listende helt stille
3. Katten spinder i hendes arme
4. Trøjen er syet af bomuld
5. Blomsterne vokser i små skåle
6. De to venner deler arbejdet
7. De sidder længe i tavshed
8. Store bølger slog mod stranden
9. Den gamle mand smilede stort
10. I regnbuen ses alle farver
11. De kørte direkte til skolen
12. Maden var rig på vitaminer
13. Konen er ældre end manden
14. Penge skal sættes i banken
15. Fødselsdagen er først på tirsdag
16. Postbudet har to små børnebørn
17. Det blev en pragtfuld ferie
18. Filmen var aldrig rigtig sjov
19. Jeg samler på gamle møbler
20. Om mandagen holder jeg fri

Reliable methods for assessing speech intelligibility are essential within hearing research, audiology, and related areas. In this thesis, two tests for speech intelligibility in Danish are developed. The first test is the Conversational Language Understanding Evaluation (CLUE), which is based on the principles of the original American-English Hearing in Noise Test (HINT). The second test is a modified version where the speech material and the scoring rules have been reconsidered. The speech material for the tests is equalized using a new procedure that produces more accurately equalized sentences than was achieved using the original HINT procedure.

A method for assessing speech intelligibility that is based on the identification of the stop-consonant [t] in a short test-word is also investigated. This method has been used previously to demonstrate the existence of an *extrinsic compensation mechanism* for reverberation in the human auditory system. However, in the present study, it is shown that the listener's perception of the test-word is not only related to the reverberation of the presented speech stimulus but also to other of its acoustic-phonetic properties. The evidence of the extrinsic compensation mechanism is therefore questionable.

Overall, the results from the present study may contribute to the development of future speech intelligibility tests in Danish and other languages. The two developed tests are expected to be useful for assessing speech intelligibility with Danish NH and HI listeners.

DTU Electrical Engineering
Department of Electrical Engineering

Ørsteds Plads
Building 348
DK-2800 Kgs. Lyngby
Denmark
Tel: (+45) 45 25 38 00
Fax: (+45) 45 93 16 34
www.elektro.dtu.dk

ISBN 978-87-92465-07-8